THE UNIVERSITY OF WARWICK

Library Declaration and Deposit Agreement

1. **STUDENT DETAILS**

Please complete the following:

Full name: …Francisco Javier Rubio………………………………………………………….

University ID number: …0866178…………………………………………………………

2. **THESIS DEPOSIT**

2.1 I understand that under my registration at the University, I am required to deposit my thesis with the University in BOTH hard copy and in digital format. The digital version should normally be saved as a single pdf file.

2.2 The hard copy will be housed in the University Library. The digital version will be deposited in the University's Institutional Repository (WRAP). Unless otherwise indicated (see 2.3 below) this will be made openly accessible on the Internet and will be supplied to the British Library to be made available online via its Electronic Theses Online Service (EThOS) service.
[At present, theses submitted for a Master's degree by Research (MA, MSc, LLM, MS or MMedSci) are not being deposited in WRAP and not being made available via EthOS. This may change in future.]

2.3 In exceptional circumstances, the Chair of the Board of Graduate Studies may grant permission for an embargo to be placed on public access to the hard copy thesis for a limited period. It is also possible to apply separately for an embargo on the digital version. (Further information is available in the *Guide to Examinations for Higher Degrees by Research*.)

2.4 *If you are depositing a thesis for a Master's degree by Research, please complete section (a) below. For all other research degrees, please complete both sections (a) and (b) below:*

(a) Hard Copy

I hereby deposit a hard copy of my thesis in the University Library to be made publicly available to readers (please delete as appropriate) EITHER immediately OR after an embargo period of ……….................... months/years as agreed by the Chair of the Board of Graduate Studies.

I agree that my thesis may be photocopied. ✓ YES / ~~NO~~ (*Please delete as appropriate*)

(b) Digital Copy

I hereby deposit a digital copy of my thesis to be held in WRAP and made available via EThOS.

Please choose one of the following options:

EITHER My thesis can be made publicly available online. ✓ YES / ~~NO~~ (*Please delete as appropriate*)

OR My thesis can be made publicly available only after…..[date] (Please give date)
YES / NO (*Please delete as appropriate*)

OR My full thesis cannot be made publicly available online but I am submitting a separately identified additional, abridged version that can be made available online.
YES / NO (*Please delete as appropriate*)

OR My thesis cannot be made publicly available online. YES / NO (*Please delete as appropriate*)

3. **GRANTING OF NON-EXCLUSIVE RIGHTS**

Whether I deposit my Work personally or through an assistant or other agent, I agree to the following:

Rights granted to the University of Warwick and the British Library and the user of the thesis through this agreement are non-exclusive. I retain all rights in the thesis in its present version or future versions. I agree that the institutional repository administrators and the British Library or their agents may, without changing content, digitise and migrate the thesis to any medium or format for the purpose of future preservation and accessibility.

4. **DECLARATIONS**

(a) I DECLARE THAT:

- I am the author and owner of the copyright in the thesis and/or I have the authority of the authors and owners of the copyright in the thesis to make this agreement. Reproduction of any part of this thesis for teaching or in academic or other forms of publication is subject to the normal limitations on the use of copyrighted materials and to the proper and full acknowledgement of its source.

- The digital version of the thesis I am supplying is the same version as the final, hard-bound copy submitted in completion of my degree, once any minor corrections have been completed.

- I have exercised reasonable care to ensure that the thesis is original, and does not to the best of my knowledge break any UK law or other Intellectual Property Right, or contain any confidential material.

- I understand that, through the medium of the Internet, files will be available to automated agents, and may be searched and copied by, for example, text mining and plagiarism detection software.

(b) IF I HAVE AGREED (in Section 2 above) TO MAKE MY THESIS PUBLICLY AVAILABLE DIGITALLY, I ALSO DECLARE THAT:

- I grant the University of Warwick and the British Library a licence to make available on the Internet the thesis in digitised format through the Institutional Repository and through the British Library via the EThOS service.

- If my thesis does include any substantial subsidiary material owned by third-party copyright holders, I have sought and obtained permission to include it in any version of my thesis available in digital format and that this permission encompasses the rights that I have granted to the University of Warwick and to the British Library.

5. **LEGAL INFRINGEMENTS**

I understand that neither the University of Warwick nor the British Library have any obligation to take legal action on behalf of myself, or other rights holders, in the event of infringement of intellectual property rights, breach of contract or of any other right, in the thesis.

Please sign this agreement and return it to the Graduate School Office when you submit your thesis.

Student's signature: ...[redacted]........ Date: ....05/June/2013....

AUTHOR: **Francisco Javier Rubio** DEGREE: **Ph.D.**

TITLE: **Modelling of Kurtosis and Skewness: Bayesian Inference and Distribution Theory**

DATE OF DEPOSIT: 05/June/2013

Modelling of Kurtosis and Skewness: Bayesian Inference and Distribution Theory

by

Francisco Javier Rubio

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy in Statistics

Department of Statistics

June 2013

Contents

List of Tables

List of Figures

Acknowledgments

First, I would like to thank my supervisor, Professor Mark F. J. Steel, for his support, encouragement, and guidance during the last three years.

I thank the examiners, professors Jim Griffin and Jim Smith, for a careful reading of this thesis.

I thank CONACyT (The National Council for Science and Technology, México) for the financial support provided.

I am grateful to the University of Warwick for covering part of my living expenses through the program “PhD hardship and completion for REF census date” . I also thank the Department of Statistics of the University of Warwick for the financial support provided for covering part of my academic fees.

To the members of the Department of Statistics of the University of Warwick as well as fellow PhD students for their support and for providing a stimulating environment.

To my love, Dialid Santiago, for her love, company, understanding, support and for making this journey an enjoyable experience.

I am deeply grateful to my parents, Jesús Rubio and Rosario Alvarez, as well as my sister Erika Rubio for their unconditional support.

Finally, I woulk like to thank Professor Olga Julià from Universitat de Barcelona for kindly providing us with the E. Coli data used in Chapter 3.

Declarations

I declare that the contents of this thesis are based on my own research in accordance with the regulations of the University of Warwick. The work in this thesis is original, unless where indicated by references. This thesis has not been submitted for examination at any other university.

All the chapters in this thesis are joint work with my supervisor Prof. Mark F. J. Steel. The contents of Chapters 2, 3, and 5 have already been accepted for publication as follows

- Rubio, F. J. and Steel, M. F. J. (2011). Inference for grouped data with a truncated skew-Laplace distribution. *Computational Statistics and Data Analysis* 55: 3218-3231.

- Rubio, F. J. and Steel, M. F. J. (2012). On the Marshall-Olkin transformation as a skewing mechanism. *Computational Statistics and Data Analysis* 56: 2251-2257.

- Rubio, F. J. and Steel, M. F. J. (2013). Bayesian inference for $P(X < Y)$ using asymmetric dependent distributions. *Bayesian Analysis* 8: 43–62.

The content of Chapter 4 is under review in an international journal. A preprint is available in the CRiSM research report

- Rubio, F. J. and Steel, M. F. J. (2011). Inference in Two-Piece Location-Scale models with Jeffreys Priors. CRiSM working paper 11–13.

The content of Chapter 6 is in preparation for publication.

Apart from the work presented in this thesis, two publications resulting from joint works during my PhD have appeared as follows

- Murillo, A. and Rubio, F. J. (2011). A note on the infinite divisibility of a class of transformations of normal variables. *Brazilian Journal of Probability and Statistics*, forthcoming.

- Rubio, F. J. and Johansen, A. M. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, forthcoming.

Abstract

This thesis is concerned with the study of distributional and inferential aspects of some classes of flexible distributions used for modelling asymmetric data.

In the last couple of decades, a great effort has been devoted to proposing new distributions that can capture departures from normality. A popular method to obtain such distributions consists of adding parameters to a known, typically symmetric, distribution. In order to do so, several classes of parametric transformations have been employed. In Chapter 2, we analyse two families of such transformations that have recently been recommended as skewing mechanisms. We show that when they are applied to several symmetric distributions, the resulting models are not flexible enough to capture moderate or high skewness. Our aims here are to show that not every parametric transformation can be used as a skewing mechanism and to emphasise the importance of assessing the flexibility of a transformed distribution using interpretable measures of skewness.

In Chapters 3–5, we focus on the study of univariate three-parameter location-scale models, where skewness is introduced by differing scale parameters either side of the location. This class of distributions is often termed *two-piece* distributions. We first present an application of a particular distribution of this kind in the context of microbiology. There, we propose a benchmark prior using the interpretability of the parameters of such model. Motivated by the importance of noninformative priors for practitioners, we then proceed to study the use of the Jeffreys prior in two-piece models and investigate the existence of the corresponding posterior distributions. We also propose a benchmark prior structure that produces proper posteriors under mild conditions for a wide class of two-piece models. In a second application, in the context of *stress-strength models*, we explore a bivariate extension of two-piece distributions using copulas. There, we also propose Bayesian models based on the interpretation of the parameters.

In Chapter 6, we introduce a five-parameter class of distributions obtained by varying both scale and shape parameters on each side of the mode. We study several aspects of this sort of models such as: subfamilies of distributions, reparameterisations, interpretation of the parameters, and two multivariate extensions. We also propose benchmark priors and illustrate their use with real data. We compare the performance of these models against appropriate competitors using several criteria.

Abbreviations

AIC	Akaike Information Criterion
BF	Bayes Factor
BIC	Bayesian Information Criterion
CDF	Cumulative Distribution Function
DIC	Deviance Information Criterion
FIM	Fisher Information Matrix
LPS	Log-Predictive Scores
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimate
PDF	Probability Density Function

Notation

The following notation is used throughout this thesis, unless otherwise stated. In addition to their statement here, they are usually described at their first occurrence. We usually use normal font type for scalars and **bold** font type for vectors, unless otherwise stated.

$\mathbb{N}$	Set of natural numbers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+$	Set of positive real numbers
$\mathbb{R}^d$	Set of $d-$dimensional real vectors
$I_A(\cdot)$	The indicator function of a set A
$I(A)$	The indicator function of the event A
$\lvert x \rvert$	The absolute value of a number x
$\mathbb{P}(A)$	The probability of the event A
$f^{-1}(\cdot)$	The inverse of $f(\cdot)$
$\sinh(\cdot)$	Hyperbolic sine
$\cosh(\cdot)$	Hyperbolic cosine
$\Gamma(\cdot)$	The Gamma function
$\mathrm{Beta}(\cdot,\cdot)$	The Beta function
$\phi(\cdot)$	The standard normal PDF
$\Phi(\cdot)$	The standard normal CDF
$\det \boldsymbol{A}$	The determinant of the matrix $\boldsymbol{A}$

Chapter 1

Introduction

> "The universe is asymmetric and I am persuaded that life, as it is known to us, is a direct result of the asymmetry of the universe or of its indirect consequences".
> Louis Pasteur,
> *L'univers est dissymetrique*.

Normality is one of the most common assumptions in the context of statistical modelling. However, important areas such as medicine, engineering, biology, finance, among others provide numerous examples of data presenting departures from normality. The need for properly modelling this sort of data has fostered a substantial effort on the development of *flexible* distributions that can capture these departures. Departures from normality of a distribution are typically studied in terms of its shape, or more specifically in terms of its *kurtosis*, and its asymmetry (also referred to as *skewness*).

In this chapter we present definitions of kurtosis and skewness as well as a summary of some popular distributions used to model these features. We discuss some inferential properties of these models as well as two general representations that can be used to produce flexible distributions. Although we explore some multivariate scenarios, the main interest of this thesis focuses on the univariate case.

1.1 Kurtosis and Skewness

Kurtosis is a term related to the peakedness and tail-weight of a distribution while skewness is related to its asymmetry. Since they can be formalised in several ways, they are often referred as "vague concepts" (Balanda and MacGillivray, 1988). Moreover, in the multivariate case the definition of *symmetry* of a distribution is not unique (Fang et al., 1990), which complicates the study of these concepts in this scenario.

Pearson (1895), Edgeworth (1904), and Pearson (1905) represent some pioneering works that introduced these concepts in terms of quantitative moment-based measures as follows

$$\begin{aligned} \text{kurtosis} &= \frac{\mu_4}{\mu_2^{4/2}}, \\ \text{skewness} &= \frac{\mu_3}{\mu_2^{3/2}}, \end{aligned}$$

where μ_j, $j = 1, 2, \ldots$, represent the jth central moment of the distribution in question, assuming they exist. Subsequently, other measures of skewness and kurtosis appeared in the literature (see Groeneveld and Meeden, 1984 and Arnold and Groeneveld, 1995 for good surveys on this).

Another formalisation of the concepts of kurtosis and skewness in the univariate case was proposed in van Zwet (1964) who presented these as *comparative concepts* based on certain partial orders on the space of distribution functions. These ideas were later studied and extended by Oja (1981). The basic definitions in these works are briefly described below.

Definition 1 *(Oja, 1981) Let $\mathcal{C}$ be the family of distribution functions. $F, G \in \mathcal{C}$ are skewness comparable if the function $G^{-1}[F(x)]$ is either convex or concave for $x \in$ Support(F).*

It follows from this definition that not every pair of distributions are skewness comparable since this definition represents a partial order in the space of distribution functions (van Zwet, 1964). This idea was used by van Zwet (1964) and Oja (1981) to define skewness as follows.

Definition 2 *(Oja, 1981) We say that $F \in \mathcal{C}$ is not more skew to the right than $G \in \mathcal{C}$ if $G^{-1}[F(x)]$ is convex for $x \in$ Support(F).*

They also defined kurtosis for symmetric distributions in terms of a partial ordering relation.

Definition 3 *(Oja, 1981) Let $\mathcal{C}_{\mathcal{S}}$ be the family of symmetric distribution functions and μ be the location of symmetry of F. We say that $F \in \mathcal{S}$ does not have more kurtosis than $G \in \mathcal{S}$ if $G^{-1}[F(x)]$ is convex for $x > \mu$.*

Balanda and Macgillivray (1990) provide an appealing intuitive explanation that relates this definition of kurtosis with with peakedness and tail-weight as follows: "an increase in kurtosis is achieved through the location - and scale - free movement of probability mass from the *shoulders* of a distribution into its centre and tails". They also provide an analogous definition of kurtosis for non-symmetric distributions which basically consists of analysing a *symmetrised version* of the distribution of interest under an appropriate transformation.

Note that definitions 2 and 3 can be used to order some families of distributions in terms of their degree of kurtosis or skewness. In addition, if we assume that F is the normal distribution, then we can assess departures from normality, in terms of these features, of a distribution G (see Jones and Pewsey, 2009 for an example of this).

Quantifying kurtosis and skewness is an interesting and extensive problem itself. In this line, several quantitative scalar and functional measures of kurtosis and skewness have been proposed (see e.g. Groeneveld and Meeden, 1984; Balanda and Macgillivray, 1990; Arnold and Groeneveld, 1995; Groeneveld, 1998; Critchley and Jones, 2008). Although the nature of these concepts does not allow for a unique way of quantifying them, Groeneveld and Meeden (1984) propose a list of desirable technical properties of these measures which are related to Definitions 2 and 3. We believe that, in addition to these properties, it is also desirable to employ an interpretable, preferably bounded, measure that allow the user to identify when a distribution is *highly/moderately skewed* or *highly/moderately kurtotic*. This will be the motivation for the choice of the measures of kurtosis and skewness employed throughout this work.

1.2 Flexible Distributions

Although there is no unique way to produce a flexible distribution, a popular method for generating distributions that can capture departures from normality in terms of kurtosis and skewness consists of adding parameters to a known, typically symmetric, distribution. The normal distribution is naturally the first candidate to be transformed. Transformations that include a parameter that controls skewness are usually referred as *skewing mechanisms* (Ferreira and Steel, 2006; Ley and Paindaveine, 2010a), while those that add a kurtosis parameter have been called *elongations* (Fischer and Klein, 2004), due to the effect produced on the shoulders and the tails of the original distribution. A third class of transformations consists of those that include parameters that can capture kurtosis and skewness jointly.

1.2.1 Asymmetric Distributions

The work by Azzalini (1985) played an important role in the popularisation of skewed distributions. In this paper he proposed the density

$$s(x;\mu,\sigma,\lambda) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\lambda\frac{x-\mu}{\sigma}\right), \tag{1.1}$$

where ϕ and Φ are the standard normal density and distribution functions, respectively, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$ and $\lambda \in \mathbb{R}$. This distribution, termed the *skew-normal distribution*, can be seen as a transformation of the normal density and contains as particular cases the normal ($\lambda = 0$), the left half-normal ($\lambda \to -\infty$), and the right half-normal ($\lambda \to \infty$) distributions. Since the density (1.1) is asymmetric for $\lambda \neq 0$, the parameter λ is interpreted as a skewness parameter. This model quickly became popular, leading to a variety of applications as well as generalisations such as those in Arnold and Beaver (2002), Wang et al. (2004), Arellano-Valle and Azzalini (2006), among others. Wang et al. (2004) show that in general, if f is a symmetric distribution and π is a function that satisfies $0 \leq \pi(x) \leq 1$ and $\pi(-x) = 1 - \pi(x)$, then $s(x) = 2f(x)\pi(x)$ is a density function. This method is known as the *skew-symmetric construction*. Since any distribution function π satisfies the required conditions, many publications appeared subsequently proposing new distributions obtained for different choices of f and π.

Another interesting family is the class of *two–piece* distributions. These distributions have a long and peculiar history, having been proposed independently in several contexts and under different levels of generality. Fechner (1897) proposed a two-piece normal obtained by joining two complementary half-normals with different scales parameters on each side of the mode (see Mudholkar and Hutson, 2000). Subsequently the same model was termed joined half-Gaussian by Gibbons and Mylroie (1973) and two-piece normal by John (1982). For a survey on independent rediscoveries of this model see Mudholkar and Hutson (2000) and Wallis (2013). Geweke (1989) applied the same idea to the normal and the Student-t distributions, the resulting distributions were termed now "split-normal" and "split-t" distributions. Fernández and Steel (1998a) proposed independently a transformation, rather close in spirit to the aforementioned models, and noted that it can be applied to any symmetric location-scale model as follows

$$s(x;\mu,\sigma,\gamma) = \frac{2}{\sigma[\gamma+1/\gamma]}\left[f\left(\frac{x-\mu}{\sigma\gamma}\right)I_{(-\infty,\mu)}(x) + f\left(\gamma\frac{x-\mu}{\sigma}\right)I_{[\mu,\infty)}(x)\right], \tag{1.2}$$

where $f(\cdot)$ is a continuous symmetric density decreasing in $|x|$, $\mu \in \mathbb{R}$ is a location parameter, $\sigma \in \mathbb{R}_+$ is a scale parameter, and $\gamma \in \mathbb{R}_+$. This transformation was termed the "Inverse scale factors". They noted that s is unimodal, with mode at μ, $s(x;\mu,\sigma,1) = f(x;\mu,\sigma)$, and that s is asymmetric for $\gamma \neq 1$. Due to these properties, an interpretable measure of skewness for this class of distributions is that proposed by Arnold and Groeneveld (1995), which is defined as one minus two times the mass cumulated to the left of the mode. This measure applied to (1.2) leads to

$$AG(\gamma) = \frac{\gamma^2 - 1}{\gamma^2 + 1},$$

which is a function that depends only on γ. They used this property to identify γ as a skewness parameter. A similar transformation was employed for the case of $f = \phi$ (standard normal) by Mudholkar and Hutson (2000)

$$s(x; \mu, \sigma, \gamma) = \left[\phi \left(\frac{x - \mu}{\sigma(1 + \gamma)} \right) I_{(-\infty,\mu)}(x) + \phi \left(\frac{x - \mu}{\sigma(1 - \gamma)} \right) I_{[\mu,\infty)}(x) \right], \tag{1.3}$$

where now $\gamma \in (-1, 1)$. A number of similar transformations were proposed subsequently by changing the parameterisation of the factors multiplying the scale parameter on each side of the mode (Kotz et al., 2001; Zhu and Galbraith, 2010). Arellano-Valle et al. (2005) proposed a representation of this kind of distributions by defining the density

$$s(x; \mu, \sigma, \gamma) = \frac{2}{\sigma[a(\gamma) + b(\gamma)]} \left[f \left(\frac{x - \mu}{\sigma a(\gamma)} \right) I_{(-\infty,\mu)}(x) + f \left(\frac{x - \mu}{\sigma b(\gamma)} \right) I_{[\mu,\infty)}(x) \right]. \tag{1.4}$$

where $\gamma \in \Gamma \subset \mathbb{R}$, $\sigma \in \mathbb{R}_+$, and $a(\gamma) > 0$ and $b(\gamma) > 0$ are known functions. This expression was initially proposed as a class of distributions but Jones (2006) found later that this is simply a reparameterisation of the same distribution, for a fixed f. Klein and Fischer (2006b) show that the parameter γ in (1.4) can be used to order this kind of distributions in the sense of Definition 2 under certain parameterisations.

A third class of transformations that has been used to induce asymmetry consists of composing the Beta distribution with a distribution function F (Jones, 2004), leading to the density

$$s(x; \alpha, \beta) = \frac{F(x)^{\alpha-1}[1 - F(x)]^{\beta-1}}{\text{Beta}(\alpha, \beta)} f(x), \tag{1.5}$$

where f is the corresponding density of F, $\alpha \in \mathbb{R}_+$, $\beta \in \mathbb{R}_+$, and $\text{Beta}(\cdot, \cdot)$ is the Beta function. If $\alpha \neq \beta$, then the resulting density is asymmetric and the degree of asymmetry depends on how different these parameters are. The cases when F normal, logistic, hyperbolic, among others have already been studied.

The fourth and last skewing mechanism described in this section is the so called *power transformation*, which simply consists of adding parameter $\alpha \in \mathbb{R}_+$ through (Lehmann, 1953)

$$s(x; \alpha) = \alpha[F(x)]^{\alpha-1} f(x). \tag{1.6}$$

It is easy to check that $s(x; 1) = f(x)$. The distributions obtained for different choices of F, such as the logistic and the normal distributions, have already been studied (Zelterman, 1987; Gupta and Gupta, 2008).

1.2.2 Kurtotic Distributions

One of the most popular methods to add a parameter that induces kurtosis is the use of a parametric transformation of the sort (Fischer and Klein, 2004)

$$T(Z) = ZW(Z; \delta), \tag{1.7}$$

where Z is a standard normal random variable and W is typically a twice-differentiable positive even function. This idea can be generalised to the use of other symmetric random variables Z with continuous distribution (Fischer and Klein, 2004). A number of candidates for W have been proposed leading to distributions with similar properties. Some examples of this sort of transformations are listed below, in all these examples the parameter $\delta \geq 0$:

(i) H transformation (Tukey, 1960)

$$W(Z; \delta) = \exp(\delta Z^2/2).$$

(ii) K transformation (Haynes et al., 1997)

$$W(Z; \delta) = (1 + Z^2)^{\delta}.$$

(iii) E transformation (Fischer and Klein, 2004)

$$W(Z; \delta) = \exp\{\delta[\cosh(Z) - 1]\}.$$

(iv) J transformation (Fischer and Klein, 2004)

$$W(Z; \delta) = \cosh(Z)^{\delta}.$$

(v) L transformation (Klein and Fischer, 2006a)

$$W(Z;\delta) = \left(\frac{\sinh(Z)}{Z}\right)^{\delta}.$$

The distributions obtained with this kind of transformations are often called *Tukey-type distributions*. A disadvantage of these models is that the density and distribution functions cannot be written in closed form and their evaluation require the numerical calculation of the inverse of (1.7).

Other popular distributions, obtained in different contexts, that contain a kurtosis parameter are: the exponential power distribution, the Student-t distribution, the symmetric sinh-arcsinh distribution (Jones and Pewsey, 2009), the symmetric $\alpha-$stable family, the symmetric Johnson-SU distribution (Johnson, 1949), among others.

1.2.3 Skew-Kurtotic Distributions

Distributions that contain parameters that control both kurtosis and skewness have been widely studied as well. Some examples of this sort of distributions are the hyperbolic distribution (Barndorff-Nielsen, 1977) and the $\alpha-$stable family of distributions. Distributions that are obtained by means of adding kurtosis and skewness parameters to symmetric ones are the Johnson SU family (Johnson, 1949); Tukey-type distributions such as the g-and-h distribution and the LambertW distribution (Tukey, 1977; Martinez and Iglewicz, 1984; Goerg, 2011), and sinh-arcsinh distributions (Jones and Pewsey, 2009). These models are primarily, although not exclusively, obtained by transforming the normal distribution. Alternatively, distributions that can account for skewness and kurtosis can be obtained by introducing skewness into a symmetric distribution that already contains a shape parameter. Examples of this class of distributions are skew-t distributions (Fernández and Steel, 1998a; Azzalini and Capitanio, 2003; Jones and Faddy, 2003; Aas and Haff, 2006; Rosco et al., 2011), and skew-Exponential power distributions (Azzalini, 1986; Fernández et al., 1995).

1.3 Representation of Transformations

In this section we describe two general representations of transformations of distributions proposed in Ferreira and Steel (2006), Abtahi and Towhidi (2011), and Ley and Paindaveine (2010a). We focus on their connection with the transformations described in Section 1.2. These representations can be used to study general properties of certain transformations (see e.g. Murillo and Rubio, 2011) and to produce new transformations.

Throughout this section S and F denote absolutely continuous distributions with support on $\mathbb{R}$. Densities are denoted with the corresponding lowercase letters.

1.3.1 Distribution-Based Transformations

Ferreira and Steel (2006) show that for any pair of continuous distributions S and F with support on $\mathbb{R}$, there exists a distribution $P : [0,1] \to [0,1]$ such that $S(x) = P[F(x)]$. The density of S can be written as $s(x) = p[F(x)]f(x)$. This implies that the transformation from a random variable X, with distribution function F, to a random variable Y, with distribution function S, can be represented as a transformation of the corresponding distributions. This representation can be used to add parameters to F by considering parametric transformations of the sort $P(\cdot;\boldsymbol{\theta})$; where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, $p \geq 1$, is a parameter. Some examples of this sort of transformations are given below.

Example 1 *The Azzalini skew-normal can be interpreted as a distribution-based transformation of the normal density ϕ by defining $p[\Phi(x);\lambda] = 2\Phi(\lambda x)$.*

Example 2 *By taking $p[F(x);\alpha,\beta] = \dfrac{F(x)^{\alpha-1}[1-F(x)]^{\beta-1}}{Beta(\alpha,\beta)}$, we obtain the transformation (1.5). Note that $p(\cdot;\alpha,\beta)$ is simply the Beta density with shape parameters (α,β).*

Example 3 *(Marshall and Olkin, 1997) proposed a transformation to add a parameter to a distribution F, defined through the density*

$$s(x;\gamma) = \frac{\gamma f(x)}{[F(x) + \gamma(1-F(x))]^2}, \quad \gamma \in \mathbb{R}_+.$$

This transformation can be easily expressed as a distribution-based transformation by defining $p[F(x);\gamma] = \frac{\gamma}{[F(x)+\gamma(1-F(x))]^2}$. This transformation is studied in more detail in Chapter 2.

Abtahi and Towhidi (2011) proposed a multivariate extension of this representation. They show that for any pair of $d-$dimensional distributions, F_d and S_d, and a random vector $\mathbf{X} = (X_1,\dots,X_d) \sim F_d$ there exists a $d-$dimensional distribution $P_d : [0,1]^d \to [0,1]$ such that

$$S_d(x_1,\dots,x_d) = P_d[F(x_1), F(x_2|x_1),\dots,F(x_d|x_1,\dots,x_{d-1})],$$

where $F(x_1)$ is the marginal of X_1 and $F(x_j|x_1 \dots x_{j-1})$ is the conditional distribution of $X_j|X_1 = x_1, \dots, X_{j-1} = x_{j-1}$, $j = 2, \dots, d$. This representation can be used to add parameters to F as well by considering parametric transformations of the sort $P_d(\cdot; \boldsymbol{\theta})$.

1.3.2 Variable-Based Transformations

A second representation was proposed by Ley and Paindaveine (2010a) as follows. They first define the group of $d-$dimensional diffeomorphisms $\mathcal{H}_d$ satisfying the conditions

(a) $H_d : \mathbf{x} = (x_1, ..., x_d) \mapsto (H_{(1)}(\mathbf{x}), ..., H_{(d)}(\mathbf{x}))$.

(b) $H_{(j)}(\mathbf{x})$ does not depend on $x_{j+1}, ..., x_d$, for $j = 1, ..., d-1$.

(c) $h_{(j)}^{x_1,...,x_j}(x_j) := H_{(j)}(\mathbf{x})$ is, for any fixed $x_1, ..., x_{j-1}$, strictly monotone increasing (and hence, invertible) with respect to x_j.

Then, they show that for any pair (S_d, F_d) of $d-$dimensional absolutely continuous distributions with support on $\mathbb{R}^d$, there exists a diffeomorphism $H_d \in \mathcal{H}_d$ such that $S_d = F_d \circ H_d$, where $\circ$ denotes composition. They also show that if F_d is symmetric, then S_d is symmetric if and only if H_d is an odd function.

This representation can also be used to add parameters to a distribution F by considering parametric diffeomorphisms of the sort $H_d(\cdot; \boldsymbol{\theta})$; where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, $p \geq 1$, is a parameter. In the univariate case we have that $H_1(x) = F_1^{-1}[S_1(x)]$ which provides an immediate connection of this representation with Definitions 2 and 3. Some examples of this sort of transformations are given below.

Example 4 *Tukey type transformations are easily interpreted as a variable-based transformation by defining $H_1(Z) = T^{-1}(Z)$, with T as in (1.7).*

Example 5 *Another example of transformations that has an immediate connection with this representation is that used to obtain the sinh-arcsinh distribution (Jones and Pewsey, 2009). The relation is obtained by defining the parametric transformation $H_1(Z; \delta, \epsilon) = \sinh[\delta \sinh^{-1}(Z) + \epsilon]$, where $\delta \in \mathbb{R}_+$ and $\epsilon \in \mathbb{R}$. Jones and Pewsey (2009) show that (δ, ϵ) can be interpreted as kurtosis and skewness parameter in the sense of Definitions 2 and 3.*

1.4 Inference

Adding parameters to a model often leads to more complicated inferences and the loss of efficiency in the estimation (Taylor et al., 2000). For this reason, it is important to assess the inferential properties of distributions obtained by means of the transformations described in previous sections.We present a brief discussion on some key classical and Bayesian results of the models described above.

1.4.1 Some Remarks on the Classical Approach

The skew-normal (1.1), being one of the most popular skewed distributions, has been extensively studied. It has been found that this model presents some inferential issues. First, Azzalini (1985) showed that the Fisher information matrix of (μ, σ, λ) is singular for $\lambda = 0$, which precludes the use of some classical results on the asymptotic behaviour of the corresponding maximum likelihood estimators (MLE). Azzalini (1985) proposed a reparameterisation that avoids this singularity but that can be difficult to use in practice. In addition, Pewsey (2000) show that the likelihood surface contains a completely flat ridge, which complicates its numerical maximisation. Finally, Ley and Paindaveine (2010b) present a complete study of the presence of a stationary point in the profile likelihood of λ at $\lambda = 0$. Ley and Paindaveine (2010b) show that these inferential issues are not specific to the skew-normal distribution but they are also present in other skew-symmetric models (Wang et al., 2004).

The family of two–piece distributions (1.4), although continuous, is not twice-differentiable at the mode precluding the use of some classical results on the asymptotic behaviour of the corresponding MLE, since this is often one of the required *regularity conditions*. However, it has been shown, using direct proofs for some specific examples (Wallis, 2013), that some asymptotical results are still valid for some two–piece models (Arellano-Valle et al., 2005). Parameterisations that induce orthogonality of the parameters (Cox and Reid, 1987) in this sort of models were studied in Jones and Anaya-Izquierdo (2010).

In the case of the power transformation (1.6), Zelterman (1987) shows that the maximum likelihood estimators of the parameters of this model do not exist if F is a logistic distribution. The existence of the MLE for a general F seems to require a case-by-case study.

The hyperbolic distribution (Barndorff-Nielsen, 1977), although it contains interpretable parameters that control kurtosis and skewness, presents some inferential issues since the convergence of the maximum likelihood estimators is quite slow. Fonseca et al. (2012) pointed out that MLE requires samples of the order of thousands to be reliable.

Although inference on Tukey-type distributions was considered intractable in the past, recent computational advances have allowed the numerical calculation of MLE for these models (Rayner and MacGillivray, 2002; Mengersen et al., 2012). In addition, Rayner and MacGillivray (2002) pointed out that this sort of models may require samples significantly larger than 100 to produce a reliable MLE. To our knowledge, properties of these estimators have been obtained basically through simulations.

1.4.2 Benchmark Bayesian Inference

Although from a Bayesian perspective the posterior distribution is well-defined under the use of proper priors, it is still important to understand the role of the parameters in order to meaningfully choose the corresponding hyperparameters. In practice it is also of interest to obtain "benchmark" (sometimes referred as "noinformative") priors. This is also a vague concept since it can be formalised in several directions but the basic idea consists of using a function of the parameters (not necessarily proper distributions) that produce posterior inferences that are close to those obtained with the classical approach (Robert, 2007). Flexible distributions tend to have more complicated expressions for their density and distribution than those of symmetric ones which complicates the study of some noninformative priors obtained by formal rules such as the Jeffreys prior (Jeffreys, 1941, 1961) and the reference prior (Berger et al., 2009). This is, for instance, the case of the Tukey-type distributions and sinh-arcsinh distributions. This difficulty, together with some issues described below, has limited the study of noninformative priors in the context of flexible distributions.

The Jeffreys prior is defined as the square root of the determinant of the Fisher information matrix (FIM). It follows then that this is not well-defined for the Azzalini skew normal (1.1), due to the singularity of the corresponding FIM. Liseo and Loperfido (2006) obtained the following expression for the reference prior of the parameters (λ, μ, σ)

$$p(\lambda, \mu, \sigma) \propto \frac{1}{\sigma} g^{1/2}(\lambda),$$

where, in their own words, "$g^{1/2}(\lambda)$ is a complicated function of the parameter of interest λ". However, they found that this function is unimodal with mode at 0, integrable over $\mathbb{R}$, the behaviour of its tails is of order $\lambda^{-3/2}$, and that the posterior is well-defined if there are at least 2 different observations. These results were later used by Bayes and Branco (2007) to construct an approximation to $g^{1/2}(\lambda)$ using a Student-t with $1/2$ degrees of freedom. Although this is an appealing tractable approximation, the fact that the likelihood of (μ, σ, λ) is flat and the priors are heavy-tailed imply heavy tailed posteriors. This is a complication in practice since it is often necessary to employ adaptive MCMC methods to properly sample from the posterior.

In the context of two–piece distributions, Fernández and Steel (2000) adopt the product prior structure

$$p(\mu, \sigma, \gamma) \propto \frac{1}{\sigma} p(\gamma),$$

where p is a proper prior on γ. This prior structure can be seen as the product of the independence Jeffreys prior of (μ, σ) in the symmetric case multiplied by a prior on the skewness parameter γ. They showed that this prior structure produce proper posteriors for a wide range of models obtained by skewing scale mixtures of normals with the mechanism described in (1.2). By considering the one-to-one relationship of the AG measure of skewness and the parameter γ in this context, Juárez and Steel (2010) proposed a prior on γ that induces a uniform prior on the AG measure of skewness.

Fonseca et al. (2012) calculated the Jeffreys prior for the parameters of the hyperbolic distribution (Barndorff-Nielsen, 1977). They obtained a complicated expression whose evaluation requires numerical integration. This is an unpleasant property since this may significantly slow down MCMC samplers.

1.5 Contribution

My contribution to the study of flexible distributions can be summarised into three parts.

(i) The first part is to show that the Marshall-Olkin transformation (Marshall and Olkin, 1997) and the power transformation (Lehmann, 1953), opposite to the conclusions in García et al. (2010), Maiti and Dey (2012) and Gupta and Gupta (2008), cannot be generally used as skewing mechanisms. We show that these transformations produce distributions that cannot capture moderate or high skewness unless they are applied to very leptokurtic distributions. The goal of this contribution is to show that adding parameters to a distribution does not automatically make it more flexible and to emphasise the importance of using interpretable measures of skewness to assess its flexibility. Part of these results are already published in Rubio and Steel (2012).

(ii) The second contribution concerns the study of Bayesian models for two-piece distributions. Benchmark and Jeffreys-type priors are obtained for a wide range of asymmetric distributions obtained through the two-piece transformation under general parameterisations. It is shown that the Jeffreys-rule prior produces improper posterior distributions in the family of two-piece scale mixtures of normals. The parameterisations that produce proper posteriors under the use of the independence Jeffreys prior for these sampling models are characterised. A general prior structure, inspired by the independence Jeffreys prior, that produces proper posteriors in a wide range of sampling models is proposed as well. The aforementioned results are submitted for publication (Rubio and Steel, 2011b). The use of these kinds of Bayesian models in the context of microbiology and stress-strength models have given rise to two publications: Rubio and Steel (2011a) and Rubio and Steel (2013). In the latter application we also explore a bivariate extension of two-piece models using a Gaussian copula.

(iii) The third contribution is to propose a generalised two-piece transformation defined on the family of location-scale distributions containing a shape parameter. The resulting distributions contain 5 parameters that have a clear interpretation. This transformation can be seen as a generalisation of the method proposed in Zhu and Galbraith (2010) for producing a generalised asymmetric Student-t distribution. It is shown that this family of transformations contains two skewing mechanisms of a different nature. Benchmark priors are proposed for these models and the properness of the posterior is studied for the case when the proposed transformation is applied to scale mixtures of normals. Some multivariate extensions are pointed out as well. These results are in preparation for publication.

1.6 Outline

The material of this work is organised as follows. In Chapter 2 we analyse the use of the Marshall-Olkin transformation (Marshall and Olkin, 1997) and the power transformation (Gupta and Gupta, 2008) as skewing mechanisms. It is shown that when these transformations are applied to several symmetric distributions, the resulting distributions cannot capture high or moderate levels of skewness in terms of interpretable measures. The remainder of the thesis focuses on the study of distributional and inferential aspects of two–piece distributions. In Chapter 3 we present an application of the two-piece Laplace distribution in the context of microbiology. We propose benchmark priors for different sorts of truncation of this model and present conditions for the existence of the corresponding posterior distributions in the presence of censored observations. In Chapter 4 Jeffreys priors and benchmark priors for two-piece models are studied in a more general framework. The existence of the posterior distribution under the use of these priors is analysed. Chapter 5 introduces Bayesian models for estimating $\mathbb{P}(X < Y)$ under both assumptions: dependence and independence of X and Y. The marginal distributions are assumed to be skewed scale mixtures of normals and the dependence is modelled using a Gaussian copula. In Chapter 6 we propose a generalisation of the family of two-piece transformations which is obtained by varying the scale and the shape parameters on each side of the mode. We present a study of distributional properties of this transformation, some subclasses of transformations, two multivariate extensions, and some reparameterisations. We also propose a location-and-scale invariant benchmark prior that produces proper posterior distributions for sampling models obtained by applying these transformations to scale mixtures of normals.

Chapter 2

On the Marshall-Olkin Transformation as a Skewing Mechanism

> "The reason why I cannot understand Shakespeare is that I want to find symmetry in all this asymmetry".
> Ludwig Josef Johann Wittgenstein,
> *Culture and Value*.

The use of the Marshall-Olkin and the power transformations as skewing mechanisms is investigated. The distributions obtained when these transformations are applied to several classes of symmetric and unimodal distributions are analysed. It is shown that most of the resulting distributions are not flexible enough to model data presenting high or moderate skewness. The only case encountered where these transformations can be considered a useful skewing mechanism is when applied to Student-t distributions with Cauchy or even heavier tails.

2.1 Introduction

The need for modeling data presenting departures from symmetry has fostered the development of more flexible classes of distributions. A popular approach is to modify a symmetric distribution by introducing a parameter that controls skewness (Azzalini, 1985; Fernández and Steel, 1998a; Jones, 2004; Ferreira and Steel, 2006).

In the context of reliability and survival analysis, Marshall and Olkin (1997) proposed a transformation of a distribution $F(x;\theta)$ that introduces a new parameter $\gamma > 0$.

This transformation is defined through the cumulative distribution function (cdf)

$$G(x;\theta,\gamma) = \frac{F(x;\theta)}{F(x;\theta)+\gamma(1-F(x;\theta))}, \tag{2.1}$$

and assuming continuity of F throughout, the corresponding probability density function (pdf) is given by

$$g(x;\theta,\gamma) = \frac{\gamma f(x;\theta)}{[F(x;\theta)+\gamma(1-F(x;\theta))]^2}. \tag{2.2}$$

The interpretation of the parameter γ is given in Marshall and Olkin (1997) in terms of the behavior of the ratio of hazard rates of F and G. This ratio is increasing in x for $\gamma \geq 1$ and decreasing in x for $0 < \gamma \leq 1$. This transformation is then proposed for the Exponential and Weibull distribution in Marshall and Olkin (1997) in order to generate more flexible models for lifetime data. Clearly, for $\gamma = 1$, G and F coincide.

Using the fact that the distribution in (2.1) describes a wider class than the original distribution F, García et al. (2010) define a generalised normal distribution (GN) by applying this transformation to a normal distribution F. They investigate the role of γ as a skewness parameter using the standardised third central moment EM $= \mu_3/\mu_2^{3/2}$ as a skewness measure (Edgeworth, 1904). In a similar search for families of skewed distributions, George and George (2011) apply the Marshall-Olkin transformation to the characteristic function of an Esscher transformed Laplace distribution (which, interestingly, leads to a very simple two-piece distribution with inverse scale factors, used later to generate data in Section 2.3.2). Maiti and Dey (2012) propose exactly the same distribution as the GN of García et al. (2010) and call it the tilted normal distribution. However, they focus mostly on its use for modelling survival data and less on the skewness properties.

We will focus here on the use of the Marshall-Olkin transformation in (2.1) as a mechanism for inducing skewness in symmetric and unimodal distributions F which are defined over the entire real line. It is immediate from (2.2) that $g(x;\theta,\gamma) = g(-x;\theta,1/\gamma)$, which means that usual measures of skewness will change sign by inverting γ and that superficially suggests γ plays the part of a skewness parameter. Perhaps the most obvious choice for F is the normal, as explored by García et al. (2010) and Maiti and Dey (2012), and we will first investigate the wider class of Student-t distributions.

In Section 2.2 we study the tail behaviour induced by the Marshall-Olkin transformation and in the next section we define a generalised t distribution based on the transformation in (2.1). We explore the role of the parameter γ in the generalised t and the generalised normal distributions using different measures of skewness and we show that the standardised third central moment can lead to counterintuitive conclusions about the shape of the density. In fact, if we use a different measure of skewness based on the relative mass both sides of the mode, it becomes clear that the Marshall-Olkin transformation applied to normal and Student-t distributions with tails that are not extremely fat is unable to accommodate even moderate amounts of skewness. Section 2.3.2 illustrates this with some simulated data. Section 2.4 examines the use of the Marshall-Olkin transformation on other classes of distributions and Section 2.5 provides some intuitive explanation of the observed behaviour. Finally, we conclude that the Marshall-Olkin transformation can not generally be used as a skewing mechanism for unimodal symmetric distributions, and we find only one exception: the Student-t distribution with Cauchy or even heavier tails.

2.2 Tail Behaviour

Marshall and Olkin (1997) proved existence of moments of (2.1) for the cases when F is Exponential or Weibull. The next Theorem shows that this transformation preserves moment existence for general F.

Theorem 1 *The moments of (2.1) exist for exactly the same order as in the original distribution F.*

Proof. *Note that if $\gamma < 1$, then*

$$\gamma < \frac{\gamma}{[F(x;\theta) + \gamma(1 - F(x;\theta))]^2} < \frac{1}{\gamma}.$$

If $\gamma > 1$, then

$$\frac{1}{\gamma} < \frac{\gamma}{[F(x;\theta) + \gamma(1 - F(x;\theta))]^2} < \gamma.$$

Therefore

$$g(x;\theta,\gamma) = K(x,\theta,\gamma)f(x;\theta),$$

where $K(x,\theta,\gamma)$ takes values in between $\min\{\gamma, 1/\gamma\}$ and $\max\{\gamma, 1/\gamma\}$. The result follows.

Theorem 1 shows that transformation (2.1) produces a distribution with exactly the same tail behaviour as the original.

2.3 Generalised t

We now define a generalised t (Gt) distribution by applying the Marshall-Olkin transformation to the Student-t distribution.

Definition 4 *A random variable X is distributed according to the generalised t distribution if its cdf and pdf are given by*

$$Gt(x;\mu,\sigma,\nu,\gamma) = \frac{F(x;\mu,\sigma,\nu)}{F(x;\mu,\sigma,\nu)+\gamma(1-F(x;\mu,\sigma,\nu))}, \tag{2.3}$$

$$gt(x;\mu,\sigma,\nu,\gamma) = \frac{\gamma f(x;\mu,\sigma,\nu)}{[F(x;\mu,\sigma,\nu)+\gamma(1-F(x;\mu,\sigma,\nu))]^2}, \tag{2.4}$$

where F and f are the cdf and pdf of a Student-t distribution with location μ, scale σ and ν degrees of freedom.

Figure 2.1 shows some examples of density (2.4) for different choices of the parameters. Of course, panel (a) is just the Student-t, whereas panel (b) corresponds to $\gamma = 0.5$ and (c) is for $\gamma = 2$. Visually, two things are worth noting about Figure 2.1: the densities generated do not seem highly skewed (even though γ is rather far from one), especially for larger values of ν, and the amount of skewness seems to depend on the value of ν. This would suggest that ν and γ can not straightforwardly be assigned roles as tail and skewness parameters, respectively.

Just as in the symmetric case, the generalised normal distribution (GN) (García et al., 2010) is a limiting case of the Gt distribution, since $\lim_{\nu\to\infty} \text{Gt}(x;\mu,\sigma,\nu,\gamma) = \text{GN}(x;\mu,\sigma,\gamma)$.

2.3.1 The Role of γ

Several measures of skewness have been proposed in the literature; see e.g. Groeneveld and Meeden (1984), Groeneveld (1991) and Arnold and Groeneveld (1995). We will assess the role of the parameter γ in the generalised t and the generalised normal by considering three different measures of skewness: the standardised third central moment EM= $\mu_3/\mu_2^{3/2}$, the Pearson measure of skewness (Pearson, 1895) defined as

$$\text{PM} = \frac{\text{Mean-Mode}}{\mu_2^{1/2}},$$

and the Arnold-Groeneveld measure of skewness (Arnold and Groeneveld, 1995), which is defined for any distribution S with unimodal density as AG $= 1 - 2S(\text{mode})$. The

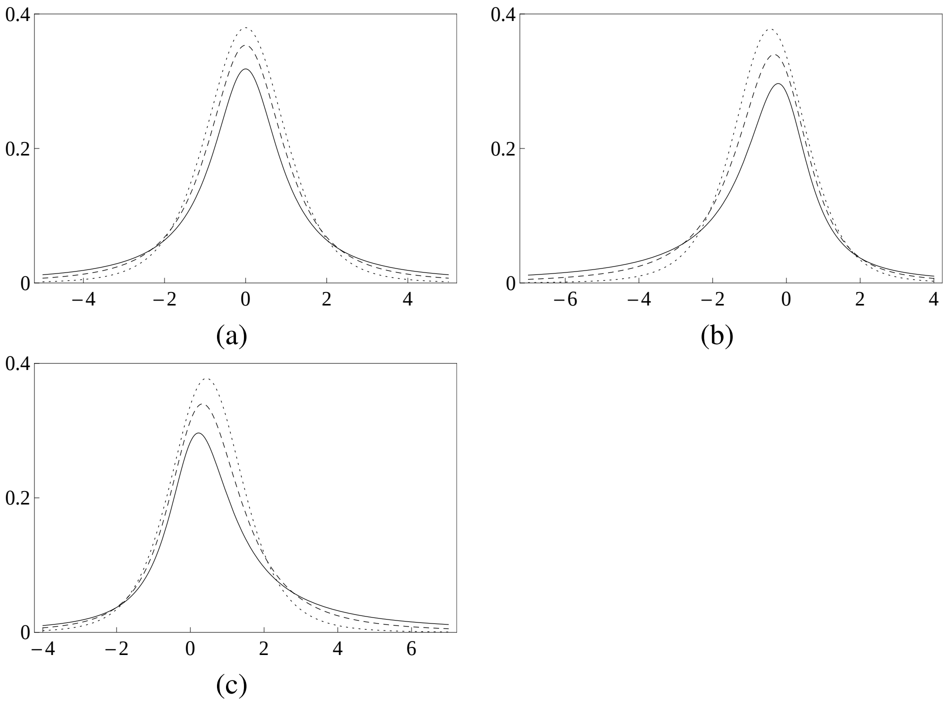


Figure 2.1: Examples of the density (2.4) with $\mu = 0$, $\sigma = 1$ and $\nu = 1$ (solid line), $\nu = 2$ (dashed line), $\nu = 5$ (dotted line): (a) $\gamma = 1$; (b) $\gamma = 0.5$; (c) $\gamma = 2$.

AG measure takes values in $(-1, 1)$, while negative values of AG are associated with left skewness and positive values of AG reflect right skewness. This skewness measure has a clear and intuitively appealing interpretation in terms of the allocation of mass both sides of the mode, and does not require the existence of any moment. We believe the three skewness measures considered here are representative of the most commonly used methods, but other measures of skewness could be used, such as the one proposed in Groeneveld and Meeden (2009) (where the direction of skewness is assumed known).

García et al. (2010) claim that the parameter γ in the generalised normal distribution plays the role of a skewness parameter as it "has a substantial effect on the skewness of the probability density function". This is shown using the standardised third central moment EM. Given that it is possible to cover a certain range of values of EM by varying the value of γ, García et al. (2010) conclude that this transformation can be used to introduce skewness. Here, however, we show that if we evaluate the role of γ using the AG and PM measures of skewness, then we have to conclude that the generalised normal and the generalised t models (except with small ν) are not flexible enough to model high or even moderate skewness.

Figure 2.2 shows the AG measure as a function of γ for several fixed values of ν

for the generalised t. While for $\nu = 1$ the behaviour seems reasonable, for larger values of ν the AG measure as a function of γ is far from surjective and not even necessarily a one-to-one function. Surprisingly, in the practically relevant case with $\nu = 10$ the parameter γ has only a very small effect on AG and the direction of this effect changes with γ. If we consider instead the moment-based measure EM in Figure 2.3, we observe a similar worrying behaviour. The parameter γ has a relatively well-defined effect on EM for small ν (of course, we need $\nu > 3$ for EM to be defined), but for larger ν the effect is very small and not monotone.

Figure 2.4 shows the AG, PM and EM measures as a function of γ for the generalised normal. This figure shows that, by varying the value of γ, the GN distribution can cover only a narrow range of values of the AG measure. For the PM and EM measures it is unclear what a reasonable range is, as they are not bounded. Thus, we will focus mostly on the AG measure in what follows. In addition, the effect of γ on AG and PM is not monotone. This clearly rules out any interpretation of γ as a skewness parameter in either the generalised normal or the generalised t models (for general ν). Interestingly, in contrast to the Student-t cases in Figure 2.3, skewness as measured by EM is positive for small γ, rather than negative. Visually, the density of the GN does not appear to have any substantial skewness even for very extreme values of γ. Clearly, EM is being driven mainly by the behaviour in the far tails.

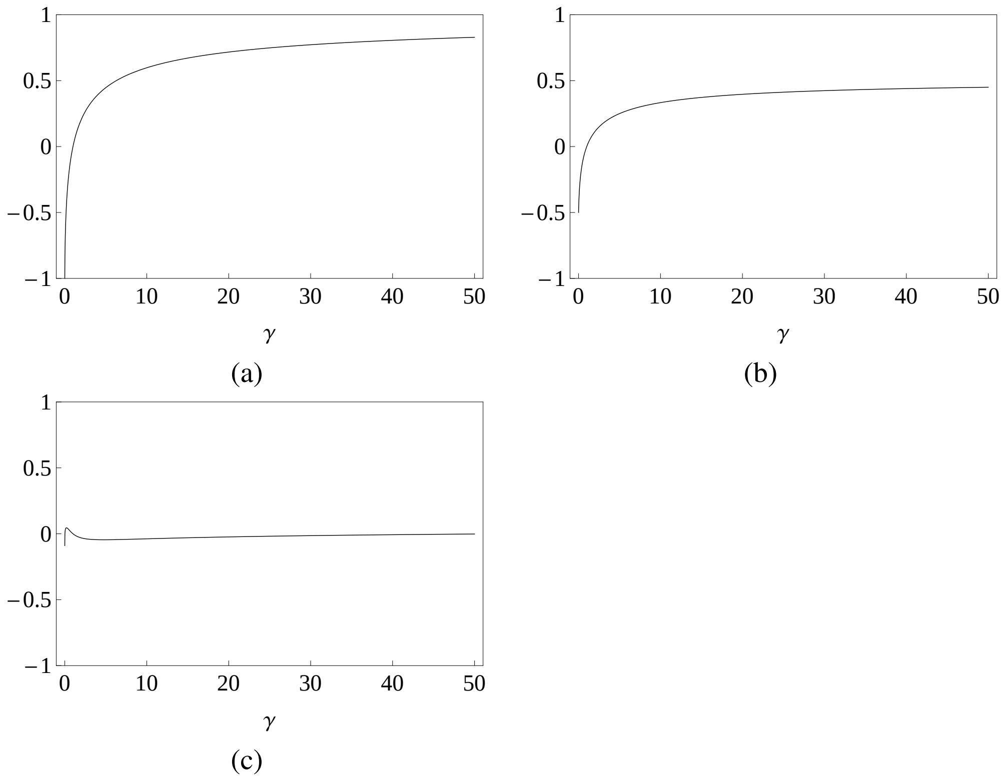


Figure 2.2: AG measure of skewness for the generalised t: (a) $\nu = 1$; (b) $\nu = 2$; (c) $\nu = 10$.

2.3.2 Example

Here we consider a simulated data set of size 500 independently sampled from a two-piece t distribution with inverse scale factors (Fernández and Steel, 1998a), which has density function

$$p(x; \mu, \sigma, \nu, \gamma) = \frac{2}{\sigma[\gamma + 1/\gamma]} \left[f(x; \mu, \sigma/\gamma, \nu) I_{(-\infty,\mu)}(x) + f(x; \mu, \sigma\gamma, \nu) I_{[\mu,\infty)}(x) \right],$$

where f is the pdf of a Student-t and we have chosen $\mu = 0$, $\sigma = 1$, $\nu = 10$ and $\gamma = 2$. The parameter $\gamma > 0$ has a clear interpretation as a skewness parameter in this model, and is linked to AG through AG$= (\gamma^2 - 1)/(\gamma^2 + 1)$. The theoretical AG measure of skewness for this example is thus 0.6.

The simulated data have somewhat heavy tails with significant right skewness. Neither the inverse scale factor transformation (see Fernández and Steel, 1998) nor the Marshall-Olkin transformation (see Theorem 1) affect the tail behaviour, so the degrees of freedom parameter ν has the same interpretation in terms of moment existence in both models. However, ν in the generalised t model affects both tail behaviour and the range

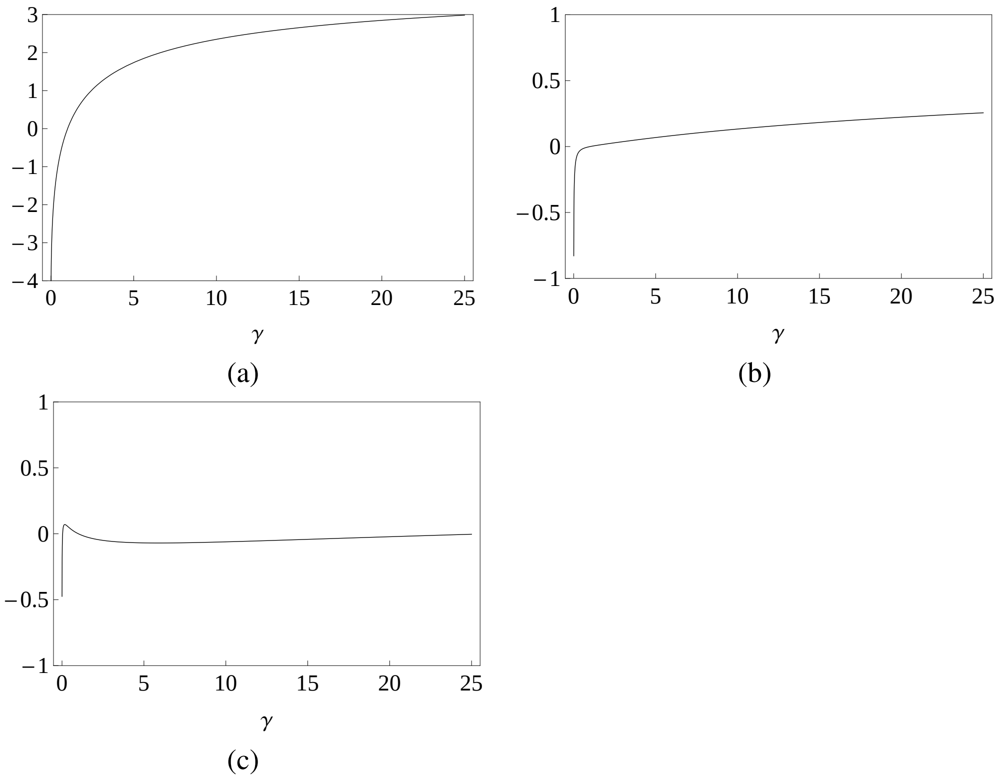


Figure 2.3: Standardised third central moment measure of skewness for the generalised t: (a) $\nu = 4$; (b) $\nu = 8$; (c) $\nu = 10$.

Model	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\nu}$	$\hat{\gamma}$	$\widehat{AG}$	AIC
Two-piece t	-0.20	0.90	11.25	2.27	0.67	1725.5
Gt	-1.88	0.06	3.58	450122.2	0.28	1739.8
GN	4.30	1.80	–	0.03	0.13	1764.2

Table 2.1: Simulated data: maximum likelihood estimates. Values for the Akaike information criterion are shown in the last column.

of possible skewness and we have seen in Figure 2.2 that the generalised t model can not account for moderate AG skewness values with $\nu = 10$. This will affect the estimation of ν and lead to a compromise estimate which is too small for the tails, but allows for some of the skewness; this produces a poor fit in the right tail and underestimation of the AG. Table 2.1 shows the maximum likelihood estimates for the model from which we generated the data, as well as the generalised t and generalised normal models. The latter two models clearly perform worse than the "true" model in terms of the Akaike information criterion (AIC), although the generalised t does better than the generalised normal, which can only allow for a very small amount of AG skewness (see Figure 2.4). This is further illustrated

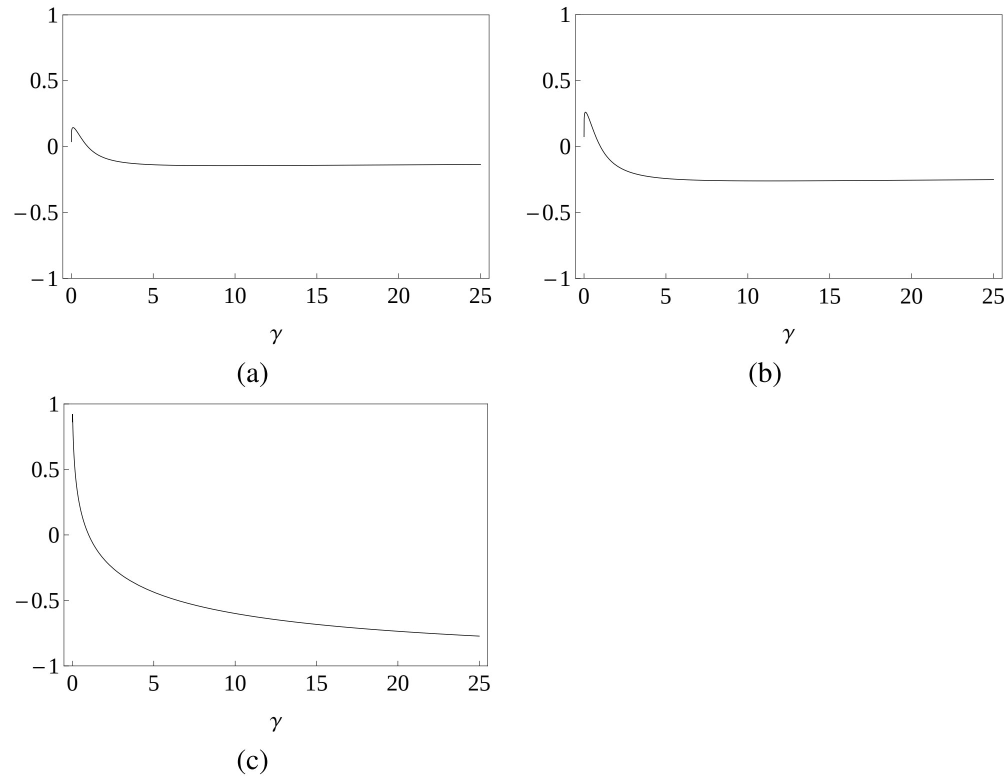


Figure 2.4: Measures of skewness of the generalised normal: (a) AG measure of skewness; (b) Pearson measure of skewness; (c) Standardised third central moment

in Figure 2.5, which presents the data histogram and the fit of the three models.

Figure 2.6 shows the estimated versus the empirical quantiles (QQ-plot), illustrating that the problem with the fit lies mainly in the right tail for the Gt and the left tail for the GN model. This example shows that the models obtained through the Marshall-Olkin transformation of normal and Student-t distributions are not flexible enough to deal with highly or moderate skewed data. Results with data simulated from a two-piece normal with the same theoretical AG skewness value similarly illustrate the lack of flexibility of the GN and Gt models.

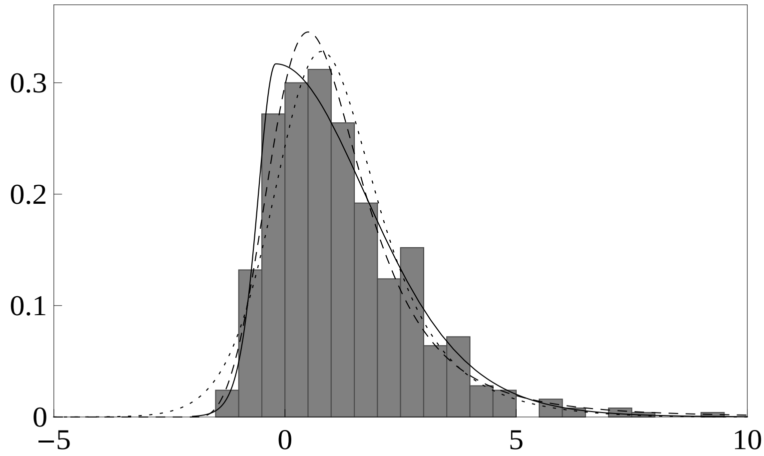


Figure 2.5: Simulated data: estimated two-piece-t density (continuous line); estimated generalised t density (dashed line); estimated generalised normal density (dotted line).

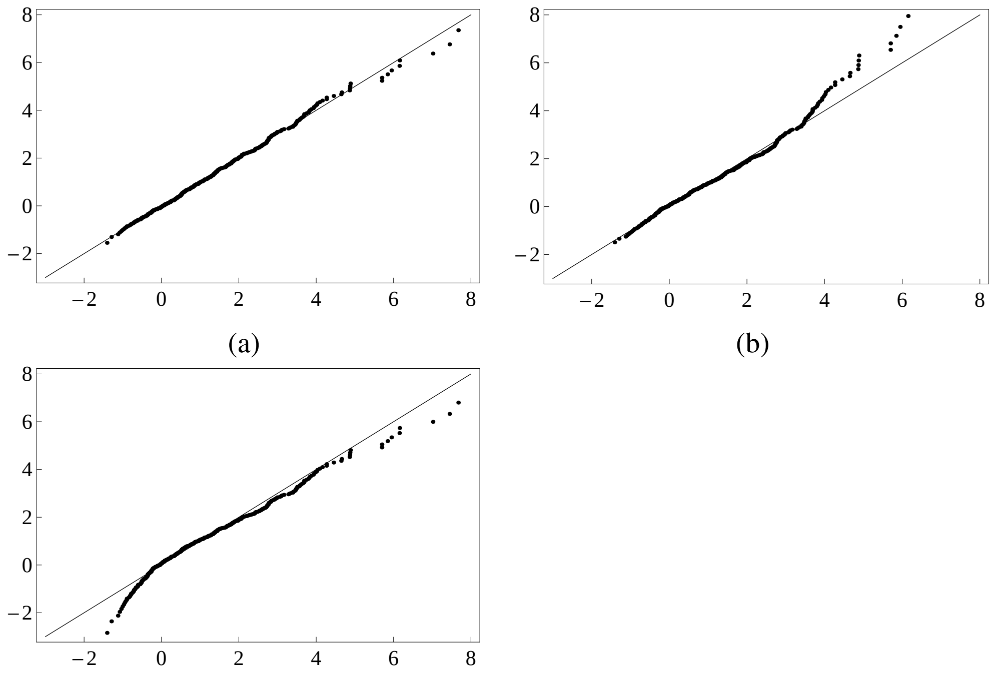


Figure 2.6: Simulated data: Estimated quantiles vs. empirical quantiles (a) Two-piece t; (b) generalised t; (c) generalised normal.

2.4 Use With Other Distributions

Let us now investigate the use of the Marshall-Olkin transformation in the context of other classes of distributions. Figure 2.7 displays the AG measure as a function of the parameter γ in (2.1) for a variety of other underlying symmetric distributions.

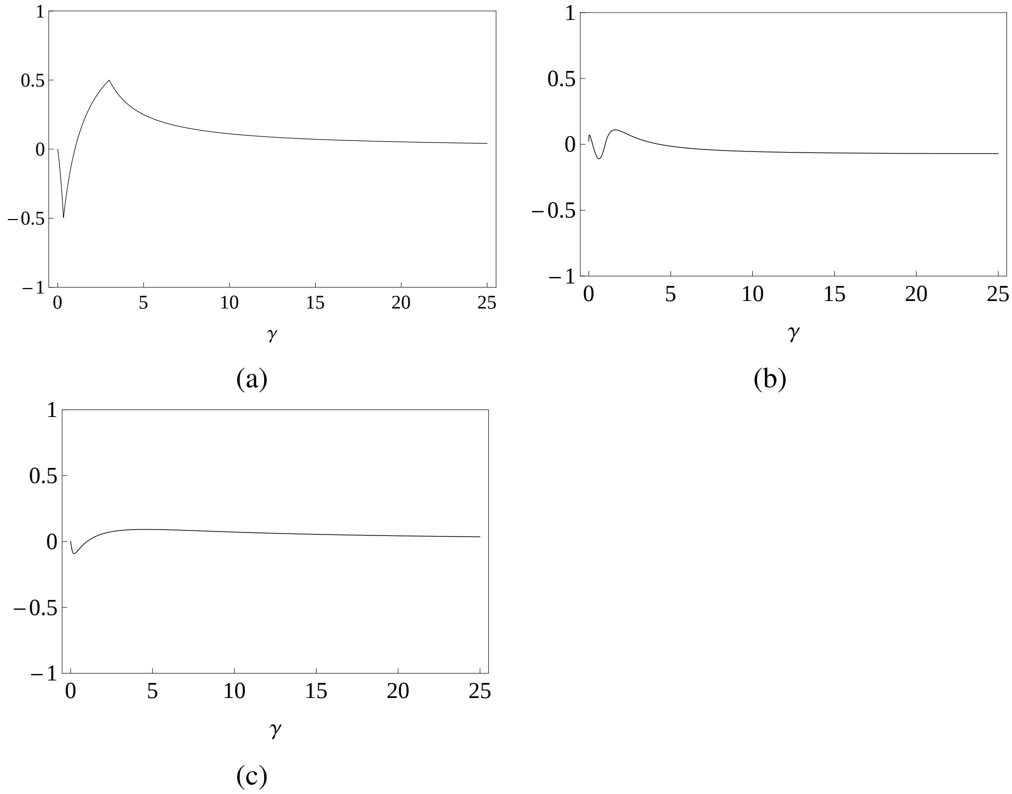


Figure 2.7: AG skewness measures as a function of γ for transformation of: (a) Laplace; (b) exponential power with $q = 3/2$; (c) hyperbolic secant distribution.

Among the choices for F we use various members of the exponential power class, which has pdf

$$f(x; \mu, \sigma, q) = \frac{1}{2^{1+(1/q)}\Gamma[1 + (1/q)]\sigma} \exp\left[-\left(\frac{|x - \mu|}{2\sigma}\right)^q\right],$$

for $q > 0$. Within this exponential power class, the Marshall-Olkin transformation produces bimodal distributions for $q < 1$ and γ sufficiently far from one, and we do not consider these distributions of practical interest for modelling. For the Laplace, which corresponds to $q = 1$, there is a single mode, which remains at zero whenever $1/3 < \gamma < 3$ and shifts to $\ln[(\gamma - 1)/2]$ for $\gamma > 3$ and to $\ln[2\gamma/(1 - \gamma)]$ for $\gamma < 1/3$. From Figure 2.7 we deduce that γ does not operate as a skewness parameter for the Laplace or the case with $q = 3/2$. Other distributions used for F are the logistic and the hyperbolic secant (Johnson et al., 1995) distributions. For the transformed logistic distribution, the AG measure is exactly zero for any value of γ. In fact, the resulting distribution is symmetric around the mode, given by $\ln(\gamma)$. The hyperbolic secant distribution is another example where we can clearly not interpret γ as a skewness parameter, as shown in Figure 2.7(c). Finally, we consider the symmetric sinh-arcsinh distribution of Jones and Pewsey (2009), which is obtained by setting their skewness parameter ϵ to zero. This distribution contains an additional parameter δ which controls the tail weight. Values of $\delta < 1$ indicate heavier tails than the normal. For small values of δ the Marshall-Olkin transformation does manage to generate substantial amounts of skewness, but for $\delta < 0.5$ the transformed density is bimodal for certain values of γ. For $\delta \geq 0.5$ the range of possible AG skewness is already quite limited and the latter is not a monotone function of γ (see Figure 2.8). Thus, none of the distributions tried in this section leads to a practically useful class of skewed distributions by using the Marshall-Olkin transformation.

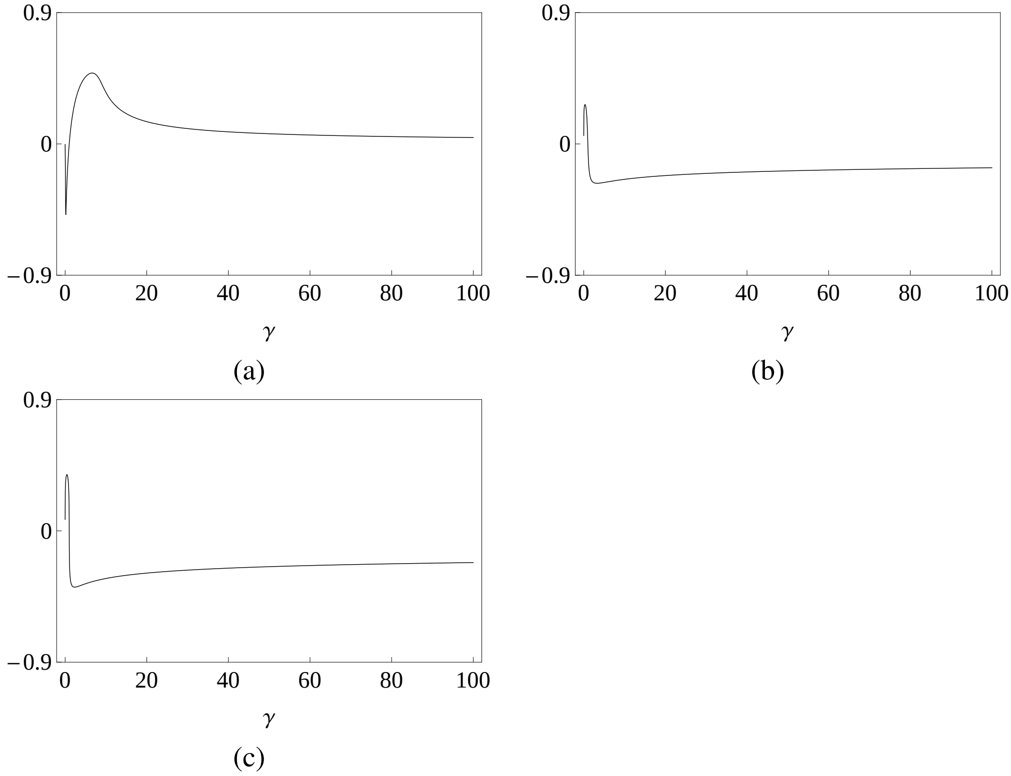


Figure 2.8: AG skewness measures as a function of γ for transformation of symmetric sinh-arcsinh distribution with: (a) $\delta = 0.5$; (b) $\delta = 1.5$; (c) $\delta = 4$.

2.5 Intuitive Explanation

Let us try to understand the effect of the Marshall-Olkin transformation on the AG measure of skewness. There are two effects going on, which can cancel each other out (and they do so exactly for the logistic). Firstly, from (2.1) it is immediate that $G(x;\theta,\gamma)$ is a decreasing function of γ for fixed (x,θ). As a consequence, if the mode would not be affected by the transformation, the AG measure would be increasing with γ. This effect is illustrated by the transformed Laplace where the mode stays at zero for $1/3<\gamma<3$, so we see in Figure 2.7(a) that AG increases with γ within this range. Secondly, however, there is the effect of a possible shift of mode. The mode is the solution of

$$f'(x;\theta)[F(x;\theta)+\gamma(1-F(x;\theta))]=2(1-\gamma)f^2(x;\theta),$$

where $f'(x;\theta)$ is the derivative with respect to x. This obviously leads to the mode of f for $\gamma=1$ and for the Logistic leads to a mode equal to $\ln(\gamma)$. If the mode (as in the latter case) increases with γ, then this second effect will make AG decrease with γ. This is illustrated again by the Laplace in Figure 2.7(a), where for γ further from one the mode shifts away from zero which quickly counteracts the first effect, making the AG value a decreasing function of γ. As γ tends to very large or very small values, the AG value tends to zero.

Generally, the behaviour of the AG skewness measure as a function of γ depends on how these two effects interact. Changes in the relative strength of these two counteracting effects also explain the lack of monotonicity we have observed for most cases.

Another way to view the way the transformation works is through the ratio $K(x,\theta,\gamma)$ between $g(x;\theta,\gamma)$ in (2.2) and the symmetric $f(x;\theta)$. Viewed as a function of x for given (θ,γ), this ratio is always in between $1/\gamma$ and γ (see also Theorem 1), but what matters most for the skewness properties of the transformation is what happens around the mode of $f(x;\theta)$. If $F(x;\theta)$ increases very slowly with x in this region, the ratio $K(x,\theta,\gamma)$ will also change slowly (whenever $\gamma\neq 1$) and the Marshall-Olkin transformation will be able to accommodate a sensible amount of skewness. If $F(x;\theta)$ is more sharply increasing, $K(x,\theta,\gamma)$ will start to behave like a step function, with the main consequence being a shift in the mode, but the distributional shape will hardly be affected. Thus, we can expect that the Marshall-Olkin transformation can only be interpreted as a skewing mechanism if it transforms extremely leptokurtic distributions (with a very small amount of mass around the mode). A complication is that for some distributions the transformation can lead to bimodality, which seriously compromises the appeal for modelling. The only example we encountered where γ can be interpreted as a skewness parameter and which avoids bimodality is the Gt with $\nu\leq 1$ degrees of freedom, explored in detail in Section 2.3.

2.6 Analysis of the Power Transformation as a Skewing Mechanism

Although we have focused on the study of a particular transformation, the main goal of this chapter is to show that adding parameters to a distribution does not automatically make it more flexible. The results presented so far suggest the need for using an interpretable measure of skewness that allows the user to identify whether or not a distribution can accommodate moderate or high skewness. In order to emphasise these points, we now investigate the use of the power transformation (Lehmann, 1953) as a skewing mechanism. This transformation has been recommended to induce skewness in several models (Nadarajah, 2006; Wagner, 2007; Gupta and Gupta, 2008). We analyse the distributions obtained by applying this transformation to several classes of unimodal symmetric distributions. It is shown that the resulting distributions can capture moderate or high skewness, in the sense of Arnold and Groeneveld (1995), only when the transformation is applied to distributions with tails heavier than those of the normal one.

Recall that the power transformation of a distribution function F is defined as

$$S(x;\alpha) = F(x)^{\alpha}, \quad \alpha \in \mathbb{R}_{+}. \tag{2.5}$$

The resulting distributions will be denoted as *power*$-F$ distributions. Figure 2.9 shows the AG measure of skewness of (2.5) for three choices of F: the normal distribution (Gupta and Gupta, 2008); the logistic distribution (Zelterman, 1987); and the hyperbolic secant distribution. Since this measure of skewness is invariant under location and scale transformations, we only consider standardised cases, this is $(\mu, \sigma) = (0, 1)$. We can observe that, by varying the parameter α, the power-normal distribution can only cover a narrow range of values of AG. On the other hand, the power-logistic and the power-hyperbolic secant distribution cover the whole range of negative values of AG. However, the positive values of AG covered by these models lie below 0.3. Figures 2.10, 2.11 and 2.12 show the AG measure of skewness as a function of α for another three choices of F: the exponential power distribution, the Student-t distribution and the symmetric sinh-arcsinh distribution (Jones and Pewsey, 2009), for several values of the corresponding shape parameters. These figures suggest that the power transformation can accommodate substantial skewness only when it is applied to very leptokurtic distributions. Therefore, the use of this transformation as a general skewing mechanism can not be recommended.

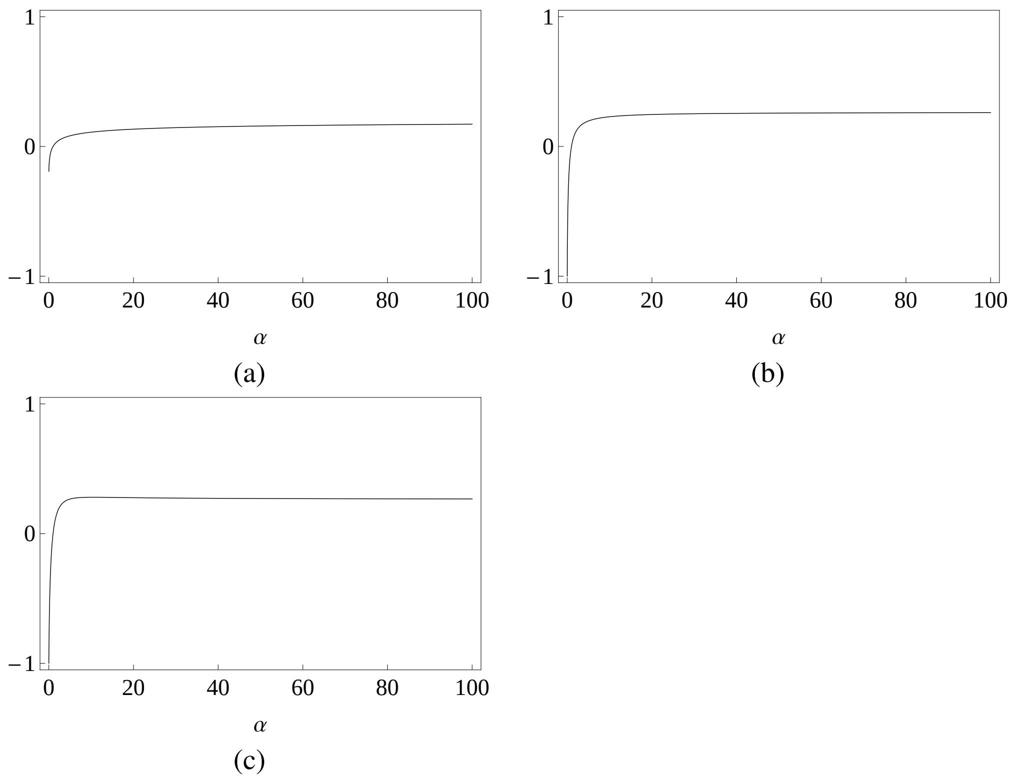


Figure 2.9: AG skewness measures as a function of α for: (a) Power-normal distribution; (b) Power-logistic distribution; (c) Power-hyperbolic secant distribution.

2.7 Conclusions

The use of the Marshall-Olkin and the power transformations as general mechanisms for inducing skewness in unimodal symmetric densities can not be recommended. They can only accommodate substantial skewness when applied to very leptokurtic distributions and can easily lead to problems of bimodality. The only case we found where they can be used in practice is when applied to a Student-t distributions with Cauchy or heavier tails. Thus, we do not recommend their use in any other situation, including the power normal of Gupta and Gupta (2008), the generalised normal of García et al. (2010) or the equivalent tilted normal of Maiti and Dey (2012). The latter case also clearly illustrates the perils of the use of the common skewness measure EM, based on the standardised third central moment. EM can be seriously misleading in practice as it can be totally dominated by the behaviour in the far tails. In addition, it is not always defined and hard to interpret as it is not bounded and is not linked to a straightforward interpretation in terms of relative mass allocation. We recommend to use and compare a range of skewness measures. The mass-based skewness measure AG is an appealing option which we find quite intuitive and interpretable.

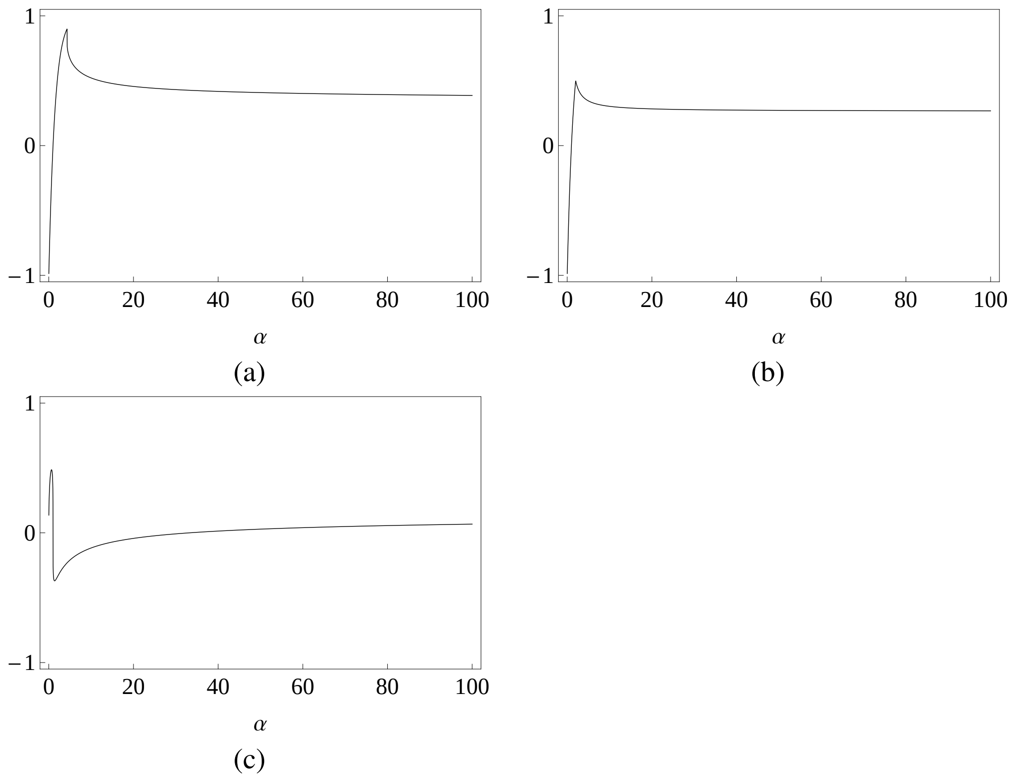


Figure 2.10: AG skewness measures as a function of α for the power-exponential power distribution with: (a) $q = 0.5$; (b) $q = 1$ (Laplace); (c) $q = 5$.

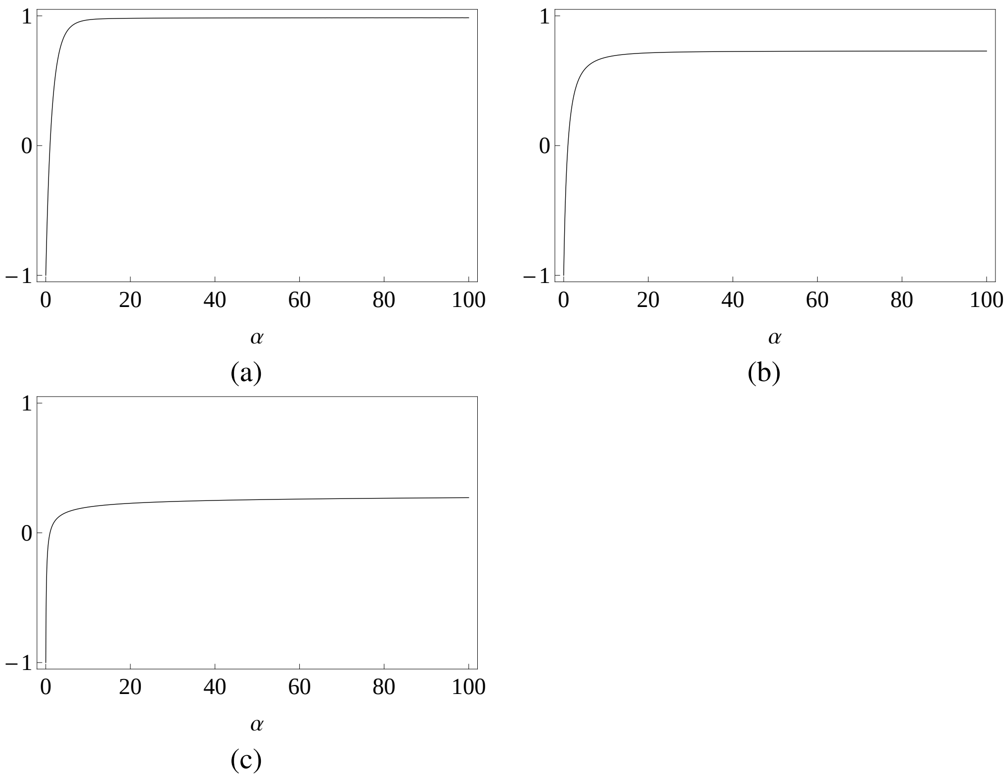


Figure 2.11: AG skewness measures as a function of α for the power-Student-t distribution with: (a) $\nu = 1$ (Cauchy); (b) $\nu = 2$; (c) $\nu = 10$.

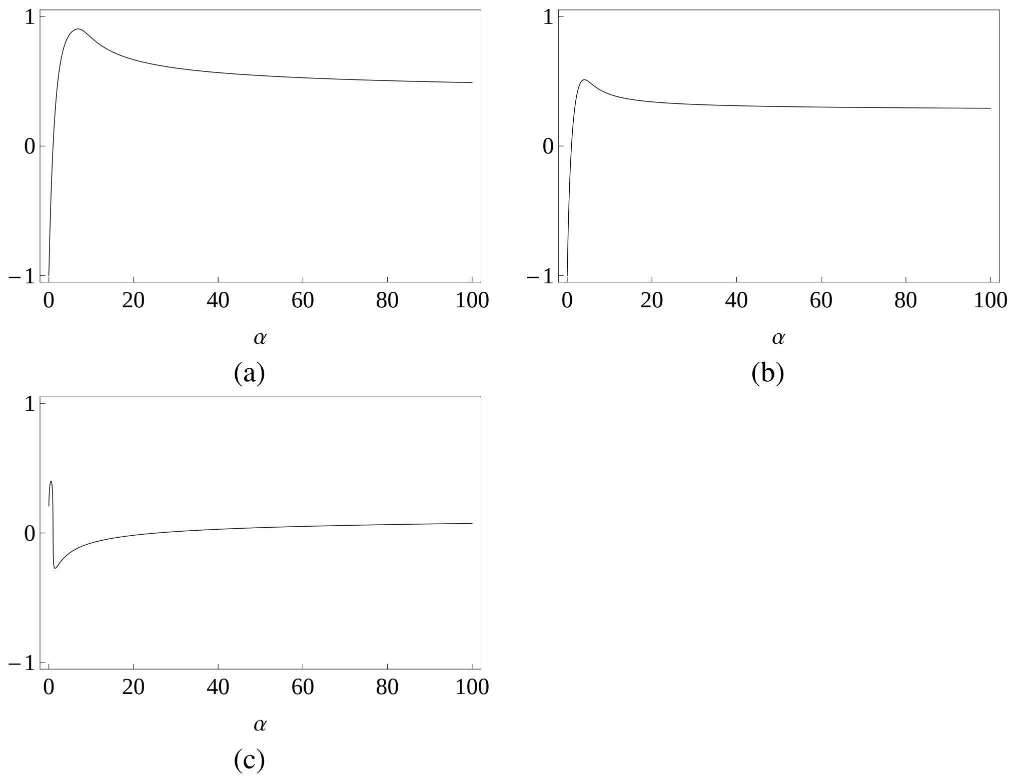


Figure 2.12: AG skewness measures as a function of α for the power-sinh-arcsinh distribution with: (a) $\delta = 0.25$; (b) $\delta = 0.5$; (c) $\delta = 4$.

Chapter 3

Inference for Grouped Data With a Truncated Skew-Laplace Distribution

> "Symmetry is only a property of dead things. Did you ever see a tree or a mountain that was symmetrical? ... Symmetry is for God, not for us."
> Louis de Bernières,
> *Captain Corelli's Mandolin.*

The skew-Laplace distribution has been used for modelling particle size with point observations. In reality, the observations are truncated and grouped (rounded). This must be formally taken into account for accurate modelling, and it is shown how this leads to convenient closed-form expressions for the likelihood in this model. In a Bayesian framework, "noninformative" benchmark priors, which only require the choice of a single scalar prior hyperparameter, are specified. Conditions for the existence of the posterior distribution are derived when rounding and various forms of truncation are considered. The main application focus is on modelling microbiological data obtained with flow cytometry. However, the model is also applied to data often used to illustrate other skewed distributions, and it is shown that our modelling compares favourably with the popular skew-Student models. Further examples on simulated data illustrate the wide applicability of the model.

3.1 Introduction

We propose a truncated skew-Laplace distribution for use with coarse (in particular rounded) or set observations. Bayesian inference will be conducted using Markov chain Monte Carlo methods. Our leading example concerns microbiological data obtained with flow cytometry, in particular forward scatter (FS) data obtained for the Escherichia Coli (E. Coli) bacterium, which are typically assumed to be proportional to bacterial size. Julià and Vives-Rego (2005, 2008) use a skew-Laplace distribution to model these data, which are truncated due to the sensitivity of the flow cytometer and are recorded as set data because the observations are presented as integers. Truncation and coarsening must be formally included in the model in order to conduct inference appropriately and to fit the data well. This application will be used throughout most of the chapter, and will serve as an important motivating example. However, later we will use the same model for a data set on the breaking strength of glass fibres, which has frequently been used in the statistics literature for illustrating skewed distributions. Further examples on simulated data illustrate the general applicability of the model.

Other distributions have, of course, been used for modelling particle size data, such as log-hyperbolic and log-Normal distributions. Fieller et al. (1992) propose a log-skew-Laplace distribution and explicitly model grouping and truncation (in line with the measurement process). Using maximum likelihood methods, they find the log-skew-Laplace to be a useful and computationally feasible alternative to log-normal and log-hyperbolic models. Instead, we use the skew-Laplace, which proves to be quite flexible and allows for modelling of observations on (any subset of) the real line while the parameters are easily interpretable. The latter greatly facilitates prior specification.

In order to define the skew-Laplace distribution, we use the general skewing framework of Fernández and Steel (1998a). This leads to a skew-Laplace distribution which is parameterised through a single skewness parameter. This skewness parameter has a nice interpretation in terms of the allocation of mass to the left and to the right of the mode. It also leads to inferential advantages as the skewing and scale parameters have clearly defined roles, which *e.g.* facilitates specification of the prior distribution. Skew-Laplace distributions with slightly different parameterisations have been used in financial modelling in Trindade and Zhu (1995) and Chen et al. (2011), using point observations.

Despite the introduction of skewness, rounding and truncation in the model, the likelihood has a relatively simple closed-form expression. This makes efficient likelihood-based inference feasible, and in this chapter we will focus on Bayesian inference. Maximum likelihood estimates, profile likelihoods and confidence intervals are numerically very close to posterior modes, posterior density functions and Highest Posterior Density (HPD) credible intervals, respectively. For models with various degrees of truncation, we propose benchmark "non-informative" priors which require the choice of a scalar prior hyperparameter. As these priors are improper, we also derive sufficient conditions for the existence of the posterior. These conditions are quite mild and trivial to check. An important advantage of the Bayesian framework is that it naturally leads to formal model comparison on the basis of Bayes factors. We compute Bayes factors between the various models as a function of the single prior hyperparameter and also consider comparison based on predictive performance. For the glass fibre data, we compare the skew-Laplace model with commonly used skew-Student specifications and find the former does better in terms of Bayes factors and matches the best skew-Student model in terms of predictive performance. Inference with the skew-Laplace model is not substantially complicated by the use of set observations or truncation of the sample space, in contrast with skew-Student or skew-normal models, for which the likelihood is not available in closed form.

3.2 Set Observations

Whenever we use a continuous model for the observations, the actually recorded values are necessarily rounded, as they are recorded to some finite precision. There has been an active literature on the quantitative effects of rounding (or grouping), as summarised in e.g. Heitjan (1989) and more recently in Schneeweiss et al. (2010). Within a Bayesian context, the explicit modelling of grouped data or set observations has been proposed by Fernández and Steel (1998b, 1999a) as a way to avoid pathological situations such as the nonexistence of a posterior with a proper prior. The reason for such behaviour is linked to the fact that any set of point observations has zero probability under a continuous sampling model. Set observations in our context of rounding are simple neighbourhoods (intervals of positive Lebesgue measure) of the recorded point observations that are chosen in accordance with the precision of the measuring process. Thus, for $i = 1, \ldots, n$ and some $d > 0$, we define

$$\mathbb{P}[\text{observing } y_j] = \mathbb{P}\left[y_j \in S_j\right] = \mathbb{P}\left[y_j - d < Y < y_j + d\right]. \tag{3.1}$$

3.3 The Skew-Laplace Distribution

In order to define the skew-Laplace distribution we use the skewness mechanism proposed in Fernández and Steel (1998a). Thus, we say that $X \sim$ skew-Laplace(μ, σ, γ) if the density function of X is

$$f_X(x; \mu, \sigma, \gamma) = \begin{cases} \frac{1}{\sigma\left(\gamma + \frac{1}{\gamma}\right)} \exp\left[\frac{\gamma(x-\mu)}{\sigma}\right] & \text{for } x < \mu\,, \\ \frac{1}{\sigma\left(\gamma + \frac{1}{\gamma}\right)} \exp\left(\frac{\mu - x}{\gamma\sigma}\right) & \text{for } x \geq \mu\,, \end{cases} \tag{3.2}$$

where $\mu \in \mathbb{R}$, $\sigma, \gamma > 0$. This model (with a different, less interpretable parameterisation, used in Julià and Vives-Rego, 2005, 2008) was called the two-piece double exponential distribution in Lingappaiah (1988). The allocation of mass to each side of the mode is given by

$$\frac{1 - F_X(\mu; \mu, \sigma, \gamma)}{F_X(\mu; \mu, \sigma, \gamma)} = \gamma^2,$$

which clearly highlights the role of γ as the skewness parameter, with μ the location parameter (which is always the mode) and σ a scale parameter. Of course, for $\gamma = 1$ we obtain the usual Laplace distribution, whereas right (positive) skewness corresponds to $\gamma > 1$ and left (negative) skewness to $\gamma < 1$. Inverting γ corresponds to mirroring the density function around the mode. If we measure skewness by the usual third centered moment divided by the cubed standard deviation, the difference between mean and mode divided by the standard deviation or the measure in Arnold and Groeneveld (1995) (defined as one minus twice the probability mass to the left of the mode), then γ and $1/\gamma$ lead to equal amounts of skewness with opposing signs. All these measures are strictly increasing functions of the skewness parameter γ.

The distribution function of X is given by

$$F_X(x; \mu, \sigma, \gamma) = \begin{cases} \frac{1}{1+\gamma^2} \exp\left(\frac{\gamma(x-\mu)}{\sigma}\right) & \text{for } x < \mu\,, \\ \frac{1}{1+\gamma^2} \left[1 - \gamma^2 \left(\exp\left(\frac{\mu - x}{\gamma\sigma}\right) - 1\right)\right] & \text{for } x \geq \mu\,. \end{cases}$$

First we investigate the analysis with the skew-Laplace distribution in (3.2), taking into account the fact that the actual observations are rounded as described in Section 3.2.

3.3.1 Likelihood Function

Consider an independent sample of rounded observations $y_1, ..., y_n$ from (3.2). The rounding as in (3.1) implies that

$$\begin{aligned} \mathbb{P}[\text{observing } y_j] &= \mathbb{P}\left[y_j - d < Y < y_j + d\right] \\ &= F_X(y_j + d; \mu, \sigma, \gamma) - F_X(y_j - d; \mu, \sigma, \gamma). \end{aligned} \tag{3.3}$$

Suppose that the sample contains k different observations $y^* = \{y_1^*, ..., y_k^*\}$ and $\{n_1, ..., n_k\}$ are the corresponding observed frequencies. The likelihood function for this sample is

$$\mathcal{L}(y; \mu, \sigma, \gamma) \propto \prod_{j=1}^{k} \left[F_X(y_j^* + d; \mu, \sigma, \gamma) - F_X(y_j^* - d; \mu, \sigma, \gamma)\right]^{n_j}.$$

The E.Coli dataset contains $n = 9,015$ observations, rounded to $k = 98$ integer values (so that $d = 1/2$), ranging from 47 to 165 with frequencies in between 1 and 306. The glass fibre data have $n = 63$ observations, rounded to the nearest one hundredth ($d = 0.005$), ranging from 0.55 to 2.24 with 49 repeated observations.

3.3.2 Bayesian Inference

In order to come up with a reasonable "noninformative" prior for the parameters in our model (3.2), we first consider the fact that the three parameters have clearly distinct roles, so that a product structure for the prior seems a good choice. In Chapter 4 we will show that the Jeffreys prior has this product structure. Then our proposal prior can also be interpreted as a slightly modified Jeffreys prior. Formal reference priors are extremely hard to derive for our analysis with set observations. If we consider the simpler case of point observations in the symmetric model (*i.e.* $\gamma = 1$) the (noninformative) full Jeffreys prior is given by $p(\mu, \sigma) \propto \sigma^{-2}$, as is the case for any location-scale model (Fernández and Steel, 1999b). However, for the skewed model the full Jeffreys and independent Jeffreys priors do not lead to a proper posterior distribution using point observations. Thus, we use a less formal approach. In particular, we modify the prior $p(\mu, \sigma) \propto \sigma^{-2}$ by bounding the parameter space of the location μ, which is important in ensuring that a posterior distribution exists (*i.e.* is a well-defined probability distribution). As we are dealing with necessarily positive observations with an internal mode in both of our applications, we use zero as a lower bound for the mode μ, whereas we introduce a single hyperparameter M as the upper bound. To elicit a prior for the skewness parameter γ, we consider the skewness measure of Arnold and Groeneveld (1995), which takes values in the interval (-1,1) and specify a uniform prior on this measure. This leads to the following prior for the model parameters:

$$\pi(\mu, \sigma, \gamma) \propto \frac{\gamma}{\sigma^2 \left(1 + \gamma^2\right)^2} I(0 < \mu \leq M). \tag{3.4}$$

Note that this density is improper in σ and the prior mass assigned to a range of positive skewness (say, $\gamma \in (a, b)$ with $b > a > 1$) is the same as that assigned to the corresponding range of negative skewness ($\gamma \in (1/b, 1/a)$). We take the upper bound M to be 1000 in the results presented in Sections 3.3-3.5.

We obtain the following sufficient condition for the existence of the posterior distribution.

Theorem 2 *The posterior distribution of* (μ, σ, γ) *for the model* (3.2) *and the prior distribution* (3.4) *is proper if the number of different observations is at least* 3*, i.e.* $k \geq 3$.
Proof. *see Appendix*

Inference for the E.Coli data was conducted using a Markov chain Monte Carlo

(MCMC) algorithm. In particular, we simulated a chain of length $2,510,000$ from the posterior using the t-walk algorithm (Christen and Fox, 2010) and after a burn-in of $10,000$ we retained every 100th set of parameter values, leading to sample of 25,000 draws. Inference with more standard MCMC methods using random walk Metropolis-Hastings steps was virtually identical. Some discussion on the t–walk algorithm seems appropriate. The t–walk is a MCMC sampler which employs two independent initial points in the parameter space Θ. The moves are proposed based on a standard Metropolis–Hastings defined on the product space $\Theta \times \Theta$, rather than simulating two independent chains. The moves are also restricted in such a manner that the resulting algorithm is invariant to scale transformations. All of these features produce an algorithm that does not require adaptive or tuning parameters. As mentioned in Christen and Fox (2010), this is an appealing characteristic of the algorithm that allows the user to focus on the data analysis rather than on the implementation of a sampling algorithm. In addition, the algorithm has already been implemented in C, Matlab, Python, and R (see Christen and Fox, 2010), which facilitates its use. This software only require the implementation of the log–posterior (up to a proportionality constant), a function to generate the two initial points, and a function that specifies the support of the posterior distribution.

Figure 3.1 shows the marginal posterior distributions of (μ, σ, γ). Inference is quite precise with 95% Highest Posterior Density (HPD) credible intervals given as follows: μ: $(69.75, 70.93)$, γ: $(1.03, 1.10)$ and σ: $(10.29, 10.73)$. It is clear that the relatively large dataset contains quite a lot of information on the three parameters in our model. The evidence indicates a relatively small but quite precisely determined amount of right skewness in the data. Prior density functions are also displayed in Figure 3.1, but they are virtually flat for the range of the parameter values shown (prior density values are quite small, so the prior for μ and γ is scaled up by the most convenient power of ten; for σ an arbitrary scaling is applied). Figure 3.2 shows the predictive distribution of the data (the sampling density in (3.2) with the parameters integrated out with the posterior distribution). However, comparing the data histogram with this predictive density indicates a rather poor fit of the data. For example, it seems that the slightly positive skewness is not consistent with the perhaps more pronounced left "shoulder" in the data when we limit ourselves to the range where data were actually observed. On the other hand, the far left tail of the predictive density is simply not matched by any data. In addition, the fit in the central part of the distribution is not very close either. Thus, it appears truncation of the data is an issue and we will now use a model that allows us to formally accommodate such truncation.

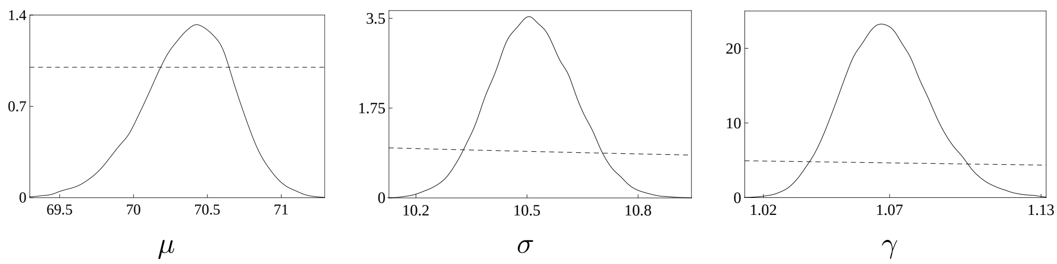


Figure 3.1: E. Coli data: Posterior (solid line) and scaled prior (dashed line) density functions.

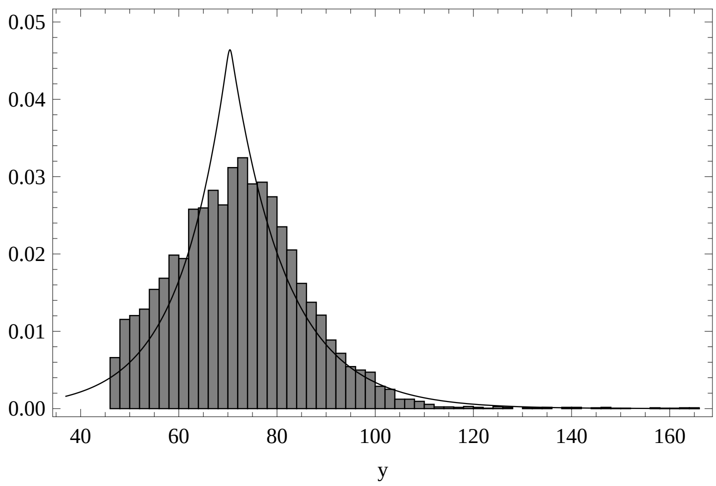


Figure 3.2: Histogram of E. Coli data and predictive density.

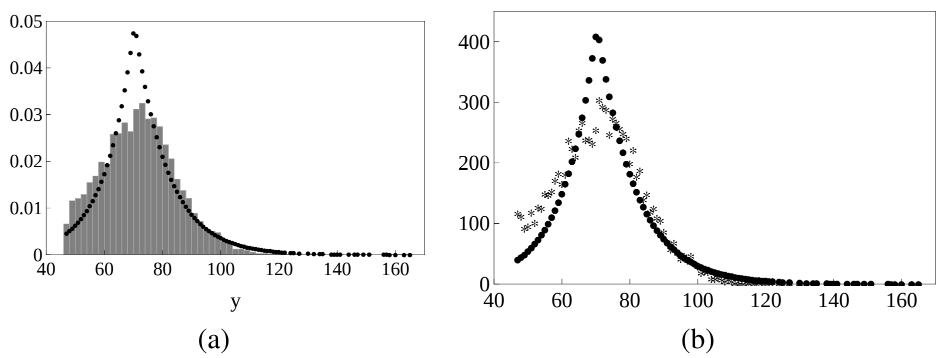


Figure 3.3: E. Coli data: (a) Normalised posterior predictive frequencies and histogram; (b) Observed frequencies (asterisks) and posterior predictive frequencies (dots).

3.4 Doubly Truncated Model

Let us consider Y to be a version of the skew-Laplace distributed random variable X in (3.2), truncated to the interval $[\theta_1, \theta_2]$. The density function of Y is then

$$f_Y(y;\mu,\sigma,\gamma,\theta_1,\theta_2) = \frac{f_X(y;\mu,\sigma,\gamma) I_{[\theta_1,\theta_2]}(y)}{F_X(\theta_2;\mu,\sigma,\gamma) - F_X(\theta_1;\mu,\sigma,\gamma)}, \tag{3.5}$$

where $\theta_1, \theta_2 \in \mathbb{R}$ and $\theta_1 < \mu < \theta_2$. Note that μ is still a location parameter (the mode), σ is a scale parameter, γ is a skewness parameter and (θ_1, θ_2) are threshold or boundary parameters. The allocation of mass to each side of the mode is given by

$$\frac{1 - F_Y(\mu;\mu,\sigma,\gamma,\theta_1,\theta_2)}{F_Y(\mu;\mu,\sigma,\gamma,\theta_1,\theta_2)} = \gamma^2 \frac{1 - \exp\left(\frac{\mu-\theta_2}{\gamma\sigma}\right)}{1 - \exp\left(\frac{\gamma(\theta_1-\mu)}{\sigma}\right)},$$

where F_Y is the distribution function of Y and is given by

$$F_Y(y;\mu,\sigma,\gamma,\theta_1,\theta_2) \quad = \quad \begin{cases} 0, & \text{for } \; y < \theta_1, \\ \frac{F_X(y;\mu,\sigma,\gamma) - F_X(\theta_1;\mu,\sigma,\gamma)}{F_X(\theta_2;\mu,\sigma,\gamma) - F_X(\theta_1;\mu,\sigma,\gamma)}, & \text{for } \; \theta_1 \le y \le \theta_2, \\ 1, & \text{for } \; y > \theta_2. \end{cases}$$

So the mass allocation both sides of the mode in this doubly truncated model is affected by γ as before but also by the boundary parameters. Of course, if $\theta_1 \to -\infty$ and $\theta_2 \to \infty$ we retrieve the previous model in the limit, but we will assume finite values for θ_1 and θ_2 in this section.

3.4.1 The Likelihood Function

An independent sample $y_1, ..., y_n$ from (3.5) rounded as in (3.1) leads to

$$\begin{aligned} \mathbb{P}[\text{observing } y_j] \quad &= \quad \mathbb{P}\left[y_j - d < Y < y_j + d\right] \\ &= \quad F_Y(y_j + d;\mu,\sigma,\gamma,\theta_1,\theta_2) - F_Y(y_j - d;\mu,\sigma,\gamma,\theta_1,\theta_2). \end{aligned}$$

As before, we suppose that the sample contains k different observations $y_1^*, ..., y_k^*$, of which the smallest is $y_{(1)}$ and the largest is $y_{(n)}$, and $n_1, ..., n_k$ are the corresponding observed frequencies. The likelihood function for this sample is

$$\begin{aligned}
\mathcal{L}(y;\mu,\sigma,\gamma,\theta_1,\theta_2) &\propto \prod_{j=1}^{k}\left[F_Y(y_j^*+d;\mu,\sigma,\gamma,\theta_1,\theta_2)-F_Y(y_j^*-d;\mu,\sigma,\gamma,\theta_1,\theta_2)\right]^{n_j} \\
&= [F_X(\theta_2;\mu,\sigma,\gamma)-F_X(\theta_1;\mu,\sigma,\gamma)]^{-n} \\
&\times\ I_{\left(-\infty,y_{(1)}-d\right]}(\theta_1)I_{\left[y_{(n)}+d,\infty\right)}(\theta_2) \\
&\times\ \prod_{j=1}^{k}\left[F_X(y_j^*+d;\mu,\sigma,\gamma)-F_X(y_j^*-d;\mu,\sigma,\gamma)\right]^{n_j}.
\end{aligned}$$

3.4.2 Bayesian Inference

Consider the following improper prior for the parameters of the sampling model in (3.5)

$$\pi(\mu,\sigma,\gamma,\theta_1,\theta_2)\propto\frac{\gamma}{\sigma^2\left(1+\gamma^2\right)^2}I(0<\theta_1<\mu<\theta_2<M), \tag{3.6}$$

which is in line with the prior (3.4) used for the untruncated model, and is again improper only in σ. Note that the prior assumes that the mode is contained within the range of observed data. This may not always seem like a reasonable assumption, but we feel that the use of a skew-Laplace model would not be natural if we were faced with data that look like one tail of such a model (we would then simply use a version of an exponential model).

The existence of the posterior is warranted by the following result:

Theorem 3 *The posterior distribution of* $(\mu,\sigma,\gamma,\theta_1,\theta_2)$ *for the Bayesian model in* (3.5) *and* (3.6) *is proper if the number of different observations is at least* 4*, i.e.* $k\geq 4$.
Proof. *See Appendix.*

We have used the same value of M and the same MCMC algorithm (with the same runlength) as in Section 3. Figure 3.4 shows the marginal prior (scaled as before) and posterior distributions for the E. Coli data.

It is interesting to note the dramatically different inference on the skewness parameter γ in this truncated model. As the data truncation is now being dealt with by the boundary parameters, we no longer need γ to reduce the mass in the left tail, and we get evidence for strong negative skewness instead, which is much more in line with the data histogram. The posterior distribution of θ_2 is flat (like the prior) over the range $(y_n+1/2, M)$ indicating that the data carry no information about θ_2 within this range. This is in line with the classical analysis, where the profile likelihood of θ_2 has an asymptote of ≈ 0.7 times the maximum value for large θ_2. There is no real data evidence to distinguish between values of θ_2 above $y_n + 1/2$ and this suggests the use of a model with only left truncation. Estimated 95%

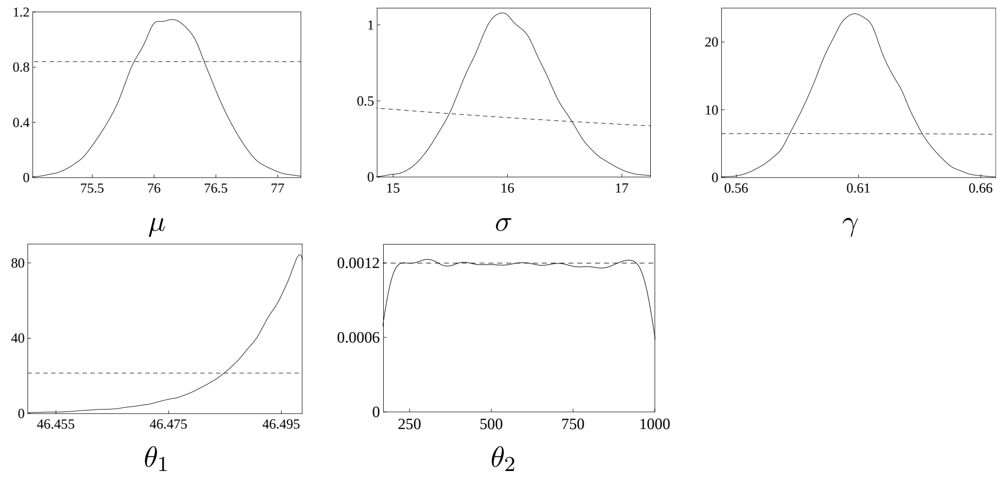


Figure 3.4: E. Coli data: Posterior (solid line) and scaled prior (dashed line) density functions.

HPD credibility intervals for the other parameters are μ: $(75.44, 76.77)$, γ: $(0.57, 0.64)$, σ: $(15.28, 16.75)$ and θ_1: $(46.47, 46.50)$.

Figure 3.5 shows the predictive density fit to the data, which is clearly much improved because of the truncation. Figure 3.6 shows the fit of the posterior predictive frequencies.

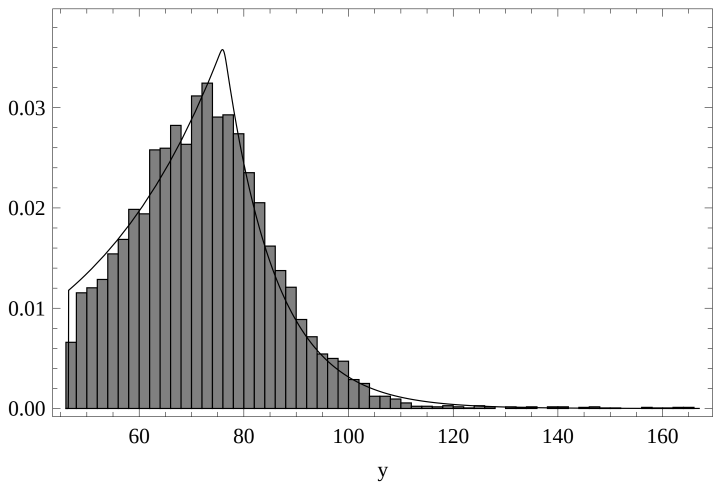


Figure 3.5: Histogram of E. Coli data and predictive density.

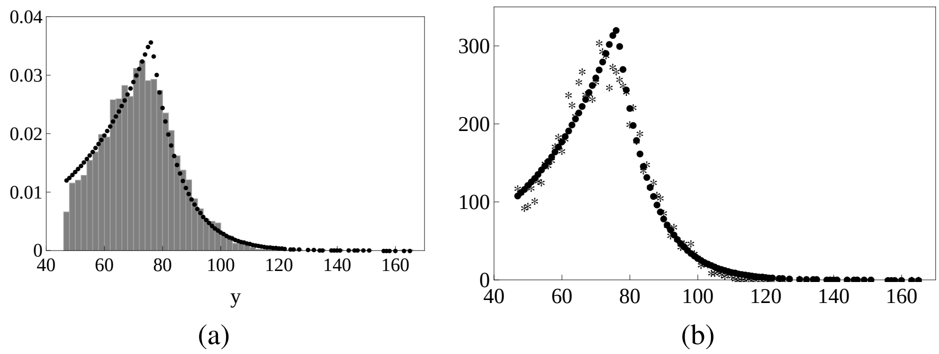


Figure 3.6: E. Coli data: (a) Normalised posterior predictive frequencies and histogram; (b) Observed frequencies (asterisks) and posterior predictive frequencies (dots).

3.5 Left Truncated Model

As the particular data used here seem to indicate that truncation on the right is superfluous, we now consider a model with only left truncation. So, let Y be a truncated version of X in $[\theta_1, \infty)$. The density function of Y is

$$f_Y(y; \mu, \sigma, \gamma, \theta_1) = \frac{f_X(y; \mu, \sigma, \gamma) I_{[\theta_1,\infty)}(y)}{1 - F_X(\theta_1; \mu, \sigma, \gamma)}. \tag{3.7}$$

Now $\theta_1 \in \mathbb{R}$ is the only threshold parameter and we restrict $\theta_1 < \mu$. The allocation of mass to each side of the mode is given by

$$\frac{1 - F_Y(\mu; \mu, \sigma, \gamma, \theta_1)}{F_Y(\mu; \mu, \sigma, \gamma, \theta_1)} = \gamma^2 \frac{1}{1 - \exp\left(\frac{\gamma(\theta_1 - \mu)}{\sigma}\right)},$$

where F_Y is the distribution function of Y and is given by

$$F_Y(y; \mu, \sigma, \gamma, \theta_1) \quad = \quad \begin{cases} 0 & \text{for } \; y < \theta_1, \\ \frac{F_X(y;\mu,\sigma,\gamma) - F_X(\theta_1;\mu,\sigma,\gamma)}{1 - F_X(\theta_1;\mu,\sigma,\gamma)}, & \text{for } \; \theta_1 \leq y. \end{cases}$$

3.5.1 The Likelihood Function

Consider an independent sample $y_1, ..., y_n$ from (3.7) with rounding as in (3.1). The likelihood function for a sample of k different observations $y_1^*, ..., y_k^*$ with frequencies $n_1, ..., n_k$ is given by

$$
\begin{aligned}
\mathcal{L}(y^*;\mu,\sigma,\gamma,\theta_1) &\propto \prod_{j=1}^{k}\left[F_Y(y_j^*+d;\mu,\sigma,\gamma,\theta_1)-F_Y(y_j^*-d;\mu,\sigma,\gamma,\theta_1)\right]^{n_j} \\
&= [1-F_X(\theta_1;\mu,\sigma,\gamma)]^{-n} I_{\left(-\infty,y_{(1)}-d\right]}(\theta_1) \\
&\times \prod_{j=1}^{k}\left[F_X(y_j^*+d;\mu,\sigma,\gamma)-F_X(y_j^*-d;\mu,\sigma,\gamma)\right]^{n_j}.
\end{aligned}
$$

3.5.2 Bayesian Inference

Consider the following improper prior for the parameters of the model (3.5)

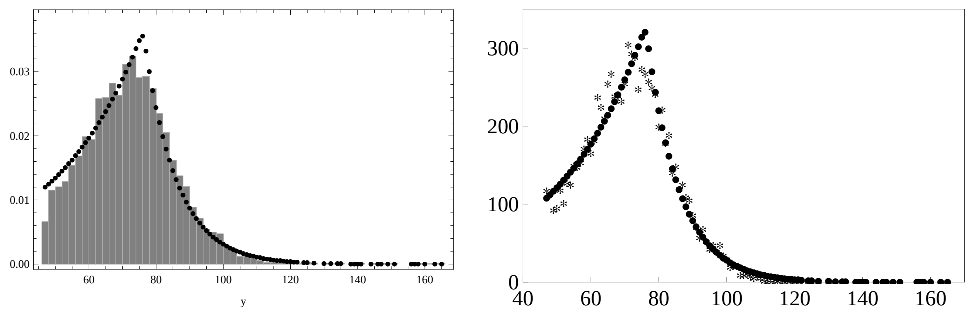
$$
\pi(\mu,\sigma,\gamma,\theta_1) \propto \frac{\gamma}{\sigma^2\left(1+\gamma^2\right)^2} I(0<\theta_1<\mu<M), \tag{3.8}
$$

which is the prior suggested by (3.6) for this reduced model.

Posterior existence is ensured by the following result:

Theorem 4 *The posterior distribution of* $(\mu,\sigma,\gamma,\theta_1)$ *for the model* (3.7) *and the prior distribution* (3.8) *is proper if the number of different observations is at least* 4*, i.e.* $k \geq 4$.
Proof. *See Appendix*

We used the same value for M and the same MCMC strategy to obtain posterior results. As expected, results are very close to the doubly truncated model, except that we do not have the right truncation parameter in the model. Marginal posterior density functions for μ, γ, σ and θ_1 are virtually identical as well as the predictive distribution. Figure 3.7 shows the fit of the posterior predictive frequencies.



(a) (b)

Figure 3.7: E. Coli data: (a) Normalised posterior predictive frequencies and histogram; (b) Observed frequencies (asterisks) and posterior predictive frequencies (dots).

3.6 Model Comparison

One advantage of Bayesian methods is that model comparison can formally be conducted by Bayes factors. The use of Bayes factors is a model selection technique which assesses the plausibility of a model $\mathcal{M}_1$ compared to that of a model $\mathcal{M}_2$ in the light of a set of observations $\mathcal{D}$. Formally, these are defined as follows (Kass and Raftery, 1995)

$$BF_{12} = \frac{\mathbb{P}(\mathcal{D}|\mathcal{M}_1)}{\mathbb{P}(\mathcal{D}|\mathcal{M}_2)}.$$

Given that this quantity can also be expressed as

$$BF_{12} = \frac{\mathbb{P}(\mathcal{M}_1|\mathcal{D})}{\mathbb{P}(\mathcal{M}_2|\mathcal{D})} \Bigg/ \frac{\mathbb{P}(\mathcal{M}_1)}{\mathbb{P}(\mathcal{M}_2)},$$

it is also interpreted as the ratio of *a posteriori* and *a priori* odds in favour of model $\mathcal{M}_1$ (Efron and Gous, 2001). Table 3.1 shows the scale proposed by Kass and Raftery (1995) for interpreting the evidence provided by this quantity.

BF_{12}	Evidence for $\mathcal{M}_1$
< 1	Negative
1–3	Barely worth mentioning
3–20	Positive
20–150	Strong
> 150	Very strong

Table 3.1: Scale of the evidence provided by the Bayes factors.

Here Bayes factors can be computed between all three models despite the arbitrary integrating constant (improperness) of the prior, since the prior has a product structure with an improper factor (in σ) which is common to all models, and the factor corresponding to model-specific parameters is integrable and thus properly normalised. The marginal likelihoods needed in the calculation of Bayes factors are estimated using importance sampling, with an importance function chosen to resemble the posterior but with fatter tails (Chopin and Robert, 2010). Some discussion on this seems necessary. Importance sampling is a stochastic numerical algorithm that can be used for estimating the integral of a function f:

$$I = \int f(x)dx,$$

The algorithm consists of rewriting I as an expectation, as follows

$$I = \int \frac{f(x)}{g(x)} g(x) dx = \mathbb{E}_g \left[\frac{f(X)}{g(X)} \right].$$

where g is a density (often termed *importance density*) function with support contained in the support of f, and X denotes a random variable with density g. Using this result, a stochastic approximation of I can be produced by obtaining a sample from g, $\{x_1, \ldots, x_N\}$, and by defining the approximation

$$I \approx \sum_{j=1}^{N} \frac{f(x_j)}{g(x_j)}.$$

This approximation is justified by the Law of Large Numbers, for a large N. Given the stochastic nature of the algorithm, a desirable property is to produce an approximation with finite variance. As pointed out by Chopin and Robert (2010), this property can be attained by choosing an importance density g with heavier tails than those of f. For example, if f is the standard normal density, then an appropriate importance function can be a standard Cauchy density.

Results with reciprocal importance sampling (Gelfand and Dey, 1994) are very close. Table 2.1 contains values for the logarithm of the Bayes factors. Information-based criteria are typically a lot easier to compute and we also present values for the BIC (Schwarz, 1978) and the DIC (Deviance Information Criterion) of Spiegelhalter et al. (2002). The BIC is defined as $-2\log(\hat{L}) + p\log(n)$, where $\hat{L}$ denotes the value of the likelihood function evaluated at the MLE, p represents the dimension of the parameter space, and n is the sample size. The model with the lower BIC among those models of interest is the one preferred. On the other hand, the DIC is defined as

$$DIC = p_D + \bar{D},$$

where $\bar{D}$ is the posterior mean deviance, the deviance is given by $D(\theta) = -2\log p(\mathcal{D}|\theta)$, $p_D = \mathbb{E}_{\theta|\mathcal{D}}[D(\theta)] + 2\log p(\mathcal{D}|\tilde{\theta}(\mathcal{D}))$, and $\tilde{\theta}(\mathcal{D})$ is typically taken as the posterior expectation of θ, $\mathbb{E}(\theta|\mathcal{D})$. In practice, the expectations involved in these expressions are usually approximated using posterior samples. Models with smaller DIC are preferable.

An alternative approach to model comparison is through the predictive performance of the models; we compute the log predictive score (LPS; see e.g. Gneiting and Raftery, 2007) based on how well the predictive distribution matches a randomly chosen prediction subsample, not used in the posterior inference. More specifically, suppose that the interest is on predicting m variables $\mathcal{D}^m = (d_1, ..., d_m)$, given the observations $\mathcal{D}$ and a model $\mathcal{M}$. The LPS criterion is then defined as minus the average of the log–predictive densities evaluated at $\mathcal{D}^m$, this is

$$LPS(\mathcal{D}^m|\mathcal{D}, \mathcal{M}) = -\frac{1}{m}\sum_{j=1}^{m} p(d_j|\mathcal{D}, \mathcal{M}),$$

where $p(\cdot|\mathcal{D}, \mathcal{M})$ denotes the predictive distribution associated to model $\mathcal{M}$. We can observe that this value is smaller when the observations $\mathcal{D}^m$ are located in regions with higher predictive density $p(\cdot|\mathcal{D}, \mathcal{M})$, which is often taken as an indicator of good predictive perfomance. For this reason, models with smaller LPS are preferable.

In practice, given that an additional sample $\mathcal{D}^m$ is not usually available, the LPS is calculated using a cross–validation technique. The idea consists of obtaining M different partitions of the original sample $\{\mathcal{D}\backslash\mathcal{D}_i, \mathcal{D}_i\}$, $i = 1, \ldots, M$, and then computing the values $LPS_i = LPS(\mathcal{D}_i|\mathcal{D}\backslash\mathcal{D}_i, \mathcal{M})$. The LPS associated to model $\mathcal{M}$ is then calculated as the average of the M values LPS_i. We use 20 prediction subsamples of 450 observations each and compute the LPS as the average over the 20 subsamples (smaller values are better).

	Model		
Criterion	untruncated	doubly trunc.	left trunc.
BIC	73020.9	71553.8	71545.4
DIC	72999.6	71516.9	71517.1
log Bayes factor	0	733	732
LPS	1822.1	1785.1	1785.8

Table 3.2: E. Coli data: Various criteria for model comparison. In the prior for the truncated models we choose $M = 1000$. Bayes factors are computed through importance sampling and we state the logarithm of the Bayes factor in favour of the model in the column versus the untruncated model. Log predictive scores (LPS) are computed on the basis of 20 partitions, each retaining 450 observations in the prediction sample.

From the results in Table 3.2 we immediately deduce that the truncated models are much preferred to the untruncated version. The relative support for both truncated models is in favour of the left truncated model if we consider BIC. The DIC, LPS and the Bayes factor all favour the doubly truncated model. Only in the case of the Bayes factor can this be interpreted in terms of posterior model probabilities: if we assume unitary prior odds, the posterior probability attached to the doubly truncated model is 2.5 times as large as that of the left truncated model. Clearly, the posterior mass assigned to the untruncated model is negligible.

Finally, we remind ourselves that the prior hyperparameter M must be selected in specifying the prior, and we know that Bayes factors can be quite sensitive to the choice of prior (Kass and Raftery, 1995). Therefore, we now investigate the sensitivity of the Bayes factor to the choice of M. Figure 3.8 shows how estimates for the marginal likelihoods and the (most relevant) Bayes factor between the truncated models vary with M. For each value of M (in the range from 200, just above the largest observation, to 2000) we run ten importance sampling estimates and the results are indicated through boxplots. Clearly, estimates are quite precise for all three models. As expected, marginal likelihood values are affected by the choice of M, since the prior domain for μ is extended beyond areas with appreciable likelihood values as M grows, so that the only real effect of larger M is that we average the likelihood with smaller prior density values, thus leading to a smaller marginal likelihood. However, the ratio of marginal likelihoods (the Bayes factor) is relatively stable as M varies. As a consequence, we consistently get slightly more support for the doubly truncated model for reasonable values of M, say $M > 300$.

3.7 Glass Fibre Data

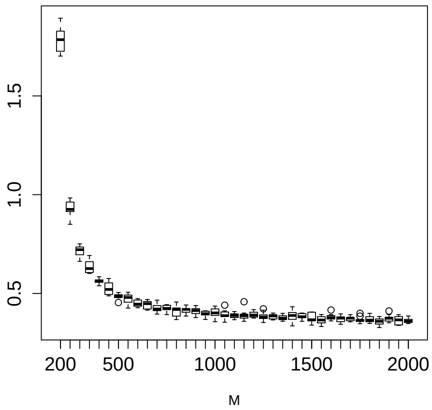
Consider the data reported in Smith and Naylor (1987) about the breaking strength of $n = 63$ glass fibres. These data were used repeatedly in the literature with a variety of skewed distributions (Jones and Faddy, 2003; Ferreira and Steel, 2006). We compare the skew-Laplace model with the more commonly used skew-Student model (with the inverse scale factor skewing of Fernández and Steel, 1998a), on the basis of set observations. This skew-Student sampling model is given by

$$f_t(x;\mu,\sigma,\gamma,\nu) = \begin{cases} \frac{2c_\nu}{\sigma\left(\gamma+\frac{1}{\gamma}\right)}\left[1+\frac{1}{\nu}\left(\frac{\gamma(x-\mu)}{\sigma}\right)^2\right]^{-(\nu+1)/2} & \text{for } x < \mu\,, \\ \frac{2c_\nu}{\sigma\left(\gamma+\frac{1}{\gamma}\right)}\left[1+\frac{1}{\nu}\left(\frac{(x-\mu)}{\gamma\sigma}\right)^2\right]^{-(\nu+1)/2} & \text{for } x \geq \mu\,, \end{cases} \tag{3.9}$$

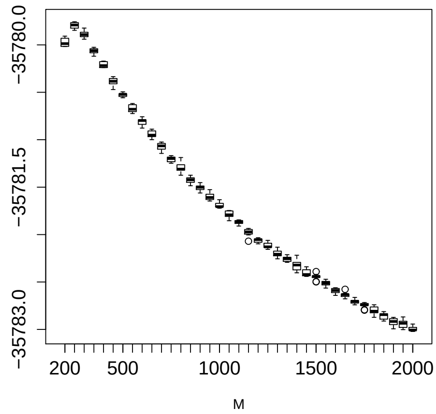
where $c_\nu = \frac{\Gamma[(\nu+1)/2]}{\Gamma[\nu/2]}\sqrt{\frac{1}{\nu\pi}}$ and $\nu > 0$ is the degrees of freedom parameter. Ferreira and Steel (2006) find that the skew-Student with $\nu = 2$ performs well for these data, but we will focus on the skew-Student with unknown degrees of freedom, as this retains the flexibility to adapt the tails to the data.

These data are breaking strengths, and therefore are subject to the physical constraint that they can not be negative. Thus, the first skew-Laplace model we consider is the left truncated one in (3.7), but with θ_1 fixed to be zero. In combination with the prior for the three model parameters in (3.4), this leads to the following result:

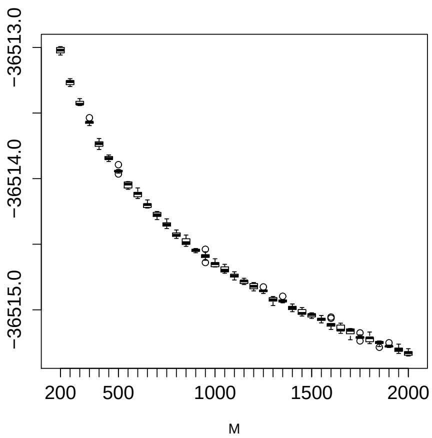
Theorem 5 *The posterior distribution of* (μ, σ, γ) *for the skew-Laplace model left trun-*



(a)



(b)



(c)

Figure 3.8: E.Coli data: Box plots based on 10 posterior samples using importance sampling. In all graphs results are given as a function of M. (a) Bayes factor in favour of the left truncated model versus the doubly truncated one. (b) Marginal likelihood for the doubly truncated model. (c) Marginal likelihood for the untruncated model.

cated at zero, i.e. (3.7) *with* $\theta_1 = 0$, *and the prior distribution* (3.4) *is proper if the number of different set-observations is at least* 4, *i.e.* $k \geq 4$.
Proof. *See Appendix*

For the skew-Student model in (3.9) we adopt the prior based on (3.4) with an extra factor for the degrees of freedom parameter

$$\pi(\mu, \sigma, \gamma, \nu) \quad \propto \quad \frac{\gamma}{\sigma^2(1+\gamma^2)^2} I(0 < \mu < M) P_\nu, \tag{3.10}$$

for which we can derive the following result on posterior existence:

Theorem 6 *The posterior distribution of* $(\mu, \sigma, \gamma, \nu)$ *for the skew-Student model in (3.9) and the prior distribution* (3.10) *is proper if the number of different set-observations is at least* 3*, i.e.* $k \geq 3$ *and* P_ν *is a proper distribution with zero mass on* $(-\infty, 1+\epsilon)$ *for any* $\epsilon > 0$.
Proof. *See Appendix*

The restriction on the prior support means that we want the predictive mean to exist, which may not be a very unreasonable assumption. Note that very small values of ν are typically associated with problems in classical likelihood inference or Bayesian inference on the basis of point observations (Fernández and Steel, 1999a). Theorem 6 also covers the case where we fix ν at any value larger than or equal to one, simply by taking P_ν to be Dirac. For the prior P_ν in the case of unknown ν we consider two possibilities: firstly, a thin-tailed gamma prior with shape parameter 2 and scale parameter 0.1, restricted to $[1+\epsilon, \infty)$ which covers a large range of values. Secondly, we adopt a hierarchical prior constructed from putting an exponential prior on the scale parameter of the gamma with shape parameter 2; this leads to the gamma-gamma prior, given by $\pi(\nu) \propto \nu/(\nu+d)^3$ with $d > 0$ and defined for $\nu \geq 1+\epsilon$. This prior has a very fat tail with no mean and shares the right-tail behaviour of the Jeffreys prior derived for the symmetric Student-t model in Fonseca et al. (2008). Here we adopt $d = 2$ which means the mode is at the boundary for $\nu = 1+\epsilon$. Throughout, we take ϵ to be machine precision (the results are the same for any $\epsilon \leq 0.0001$).

Figure 3.9 shows the inference on the parameters of the zero truncated skew-Laplace model using $M = 10$, and it is clear that skewness is again an important aspect of the data. As in other studies with this application, we find clear evidence of negative skewness.

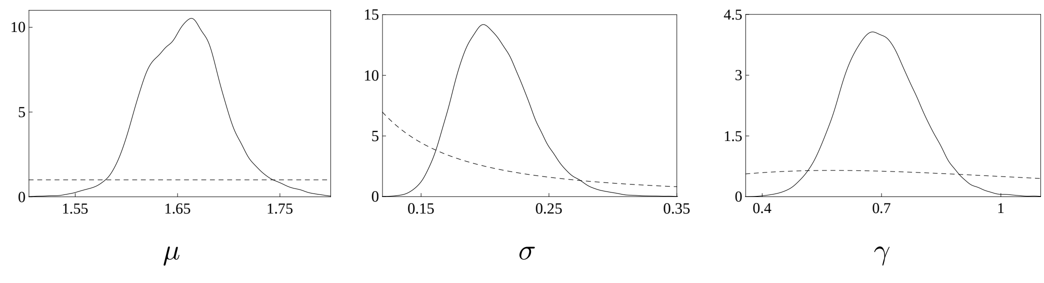


Figure 3.9: Glass data: Posterior (solid line) and prior (dashed line) density functions for the Laplace model.

Posterior predictive density functions are overplotted with a histogram (chosen according to Sturges' formula) of the data in Figure 3.10. All models seem to fit the data reasonably well, but there are some differences between the predictives. It is interesting to note that the skew-Laplace model does not lead to such a sharp peak as in the application with the E. Coli data. The fact that the data are not very peaked means there is some posterior uncertainty regarding the mode (see Figure 3.9), and this is reflected in the posterior predictive (which is simply the sampling model integrated out with the posterior). As a consequence, the skew-Laplace and the skew-Student model with $\nu = 2$ are actually very similar. Thus, the simple skew-Laplace model adapts to the data at hand.

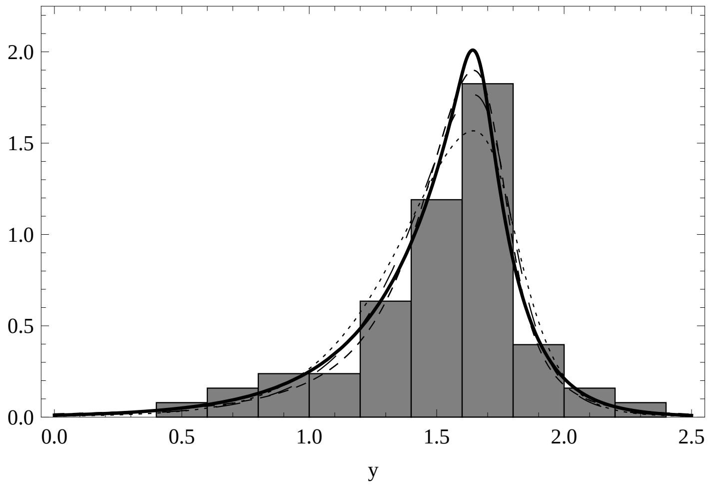


Figure 3.10: Histogram of glass data, predictive density for the skew-Laplace (bold line), skew-t_2 (short dashes), skew-t_ν with gamma prior (dotted line), skew-t_ν with gamma-gamma prior (long dashes).

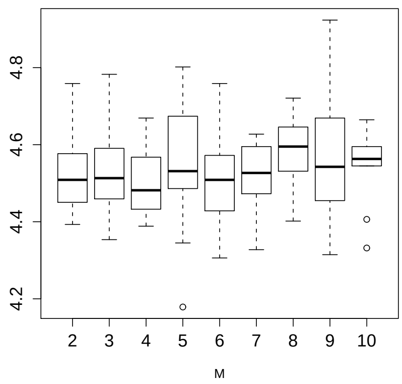
3.7.1 Model Comparison

In order to have a more formal comparison of the different models, we can again compute the Bayes factors. Marginal likelihood estimates depend on M, as discussed in Section 6, and this leads to the Bayes factors displayed in Figure 3.11. These are Bayes factors in favour of the zero truncated skew-Laplace model as a function of M and the boxplots correspond to ten importance sampling estimates.

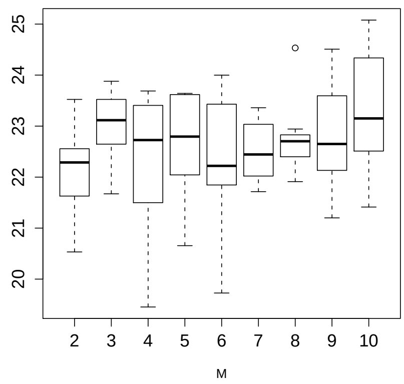
Clearly, the skew-Laplace model beats the skew-Student models. Among the skew-Student models, it seems best to fix ν to be a suitable value for these data, namely $\nu = 2$. The value of M does not seem to have a systematic effect on these Bayes factors. Of course, truncation is not built into the skew-Student models, but this aspect is not that important for the Bayes factors, as the untruncated skew-Laplace model does almost equally well with these data (*e.g.* the Bayes factor is around 1.24 in favour of the zero truncated skew-Laplace model for $M = 10$). Truncation is, however, not that easily implemented in the skew-Student models, both in terms of computational ease and proving results such as Theorem 6.

To assess the impact of the different priors on ν, we overplot posterior and prior density functions for ν in Figure 3.12. Despite its fatter right tail, the gamma-gamma prior has a mode closer to zero and leads to more posterior mass concentration on small values of ν. Thus, the predictive and the marginal likelihood are closer to that of the case with $\nu = 2$ than with the gamma prior.

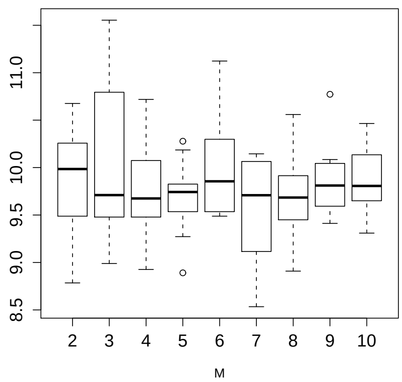
We compare the models in terms of their predictive performance by computing log predictive scores, averaged over 20 partitions of the data where 20 randomly chosen observations are used in the prediction subsample, and the results are presented in Table 3.3. The skew-Laplace and skew-t_2 models predict best and are roughly equally good.



(a)



(b)

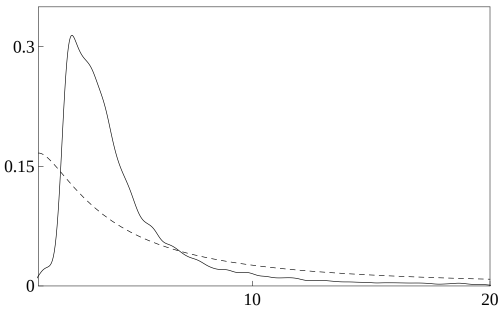


(c)

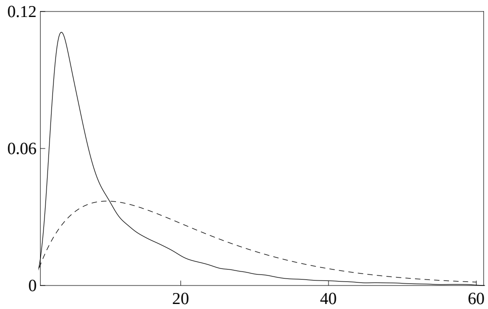
Figure 3.11: Glass data: Bayes factors as a function of M in favour of the zero truncated skew-Laplace model versus (a) skew-t_2 model; (b) skew-t_ν model with gamma prior; (c) skew-t_ν model with gamma-gamma prior

Model	Skew-Laplace	Skew-t gamma-gamma	Skew-t gamma	Skew-t_2
LPS	95.76	96.28	96.08	95.74

Table 3.3: Glass data: Log predictive scores (LPS), computed on the basis of 20 partitions, each retaining 20 observations in the prediction sample.



(a)



(b)

Figure 3.12: Glass data: degrees of freedom parameter ν for skew-Student (a) Posterior distribution of ν (solid line) and gamma-gamma prior (dashed line). (b) Posterior distribution of ν (solid line) and gamma prior (dashed line).

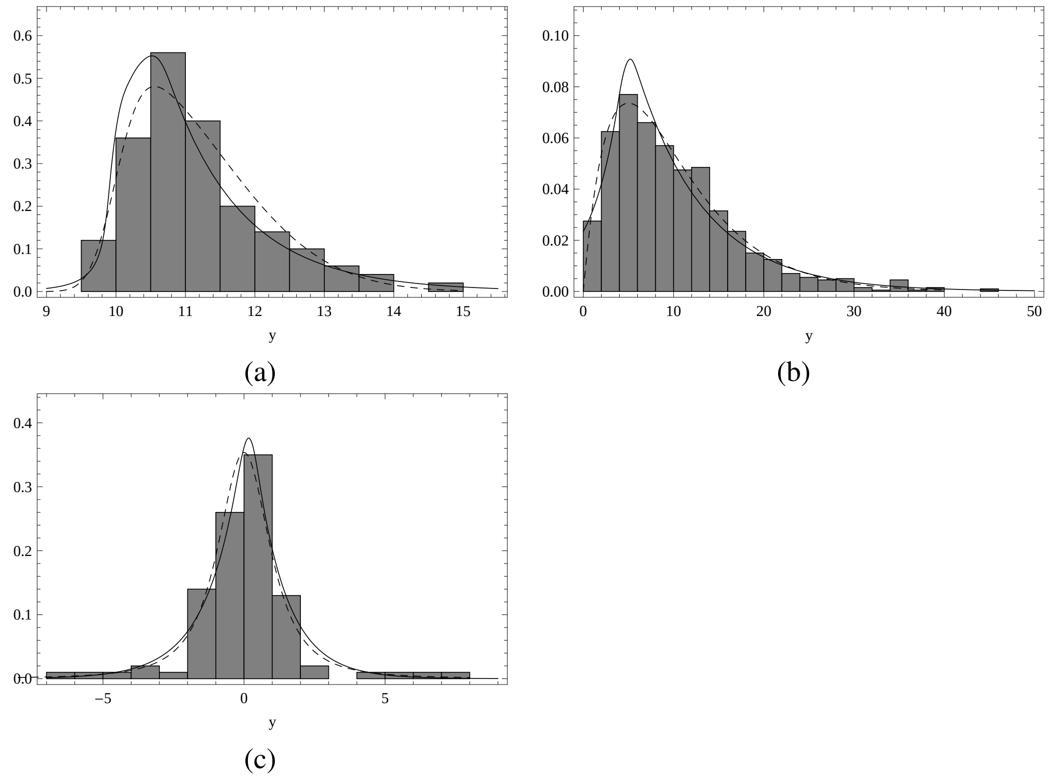
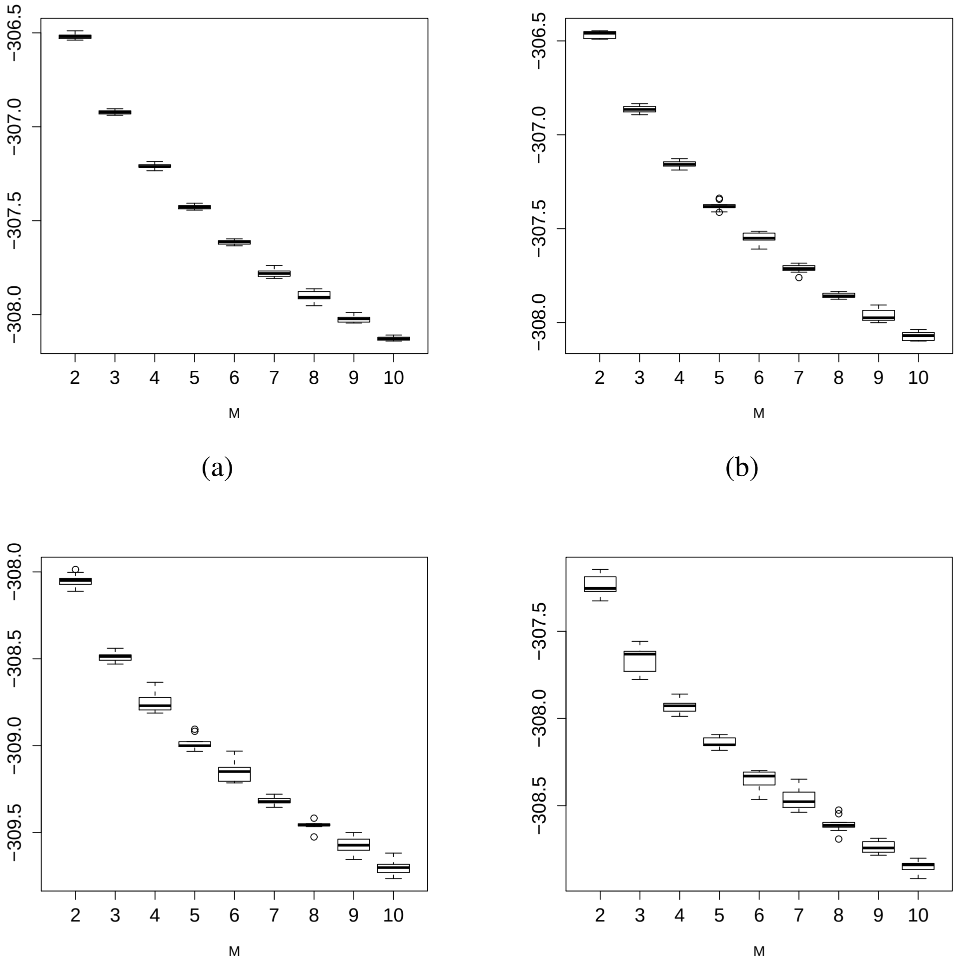


Figure 3.13: Simulated data: Skew-Laplace predictive (solid line) and data-generating density (dashed line) with data histogram in grey. Data generated from (a) Azzalini skew-normal ($n = 100$) (b) Gamma(2,5) with zero truncated skew-Laplace ($n = 1000$) ; (c) t_2 ($n = 100$).



(a) (b)

(c) (d)

Figure 3.14: Glass data: Box plots based on 10 posterior samples using importance sampling. In all graphs results are given as a function of M. (a) Marginal likelihood estimate for the Laplace model. (b) Marginal likelihood estimate for the t_2 model.(c) Marginal likelihood estimate for the t_ν model with exponential prior. (c) Marginal likelihood estimate for the t_ν model with Juárez-Steel prior.

3.8 Conclusions

In this chapter, we describe inference with the skew-Laplace model, a flexible model for use with unimodal data sets where rounding and truncation of the data are possibly important issues. We formally incorporate rounding of the data and truncation of the support in the analysis. For four versions of the model (untruncated support, finite support with unknown boundaries, left truncated support with unknown boundary, left truncated at zero), we specify a fairly noninformative and sensible prior which only depends on a single hyperparameter M and we derive sufficient conditions for the existence of the posterior. These conditions refer to the number of different observations in the sample, are trivial to check and are very likely to be satisfied in samples of practical interest. The particularly tractable nature of the skew-Laplace model makes it easy to introduce rounding and truncation, both for computational implementations and for proofs of posterior existence. In particular, the likelihood of the model is available in closed form, in contrast with many other models, such as the skew-normal or skew-Student (e.g. using the skewing ideas of Azzalini, 1985, Fernández and Steel, 1998a, or Jones and Faddy, 2003).

The skew-Laplace model behaves well in the motivating application on flow cytometry data, as could perhaps be expected. However, it also beats the skew-Student in the glass fibre data set, an application for which skew-Laplace modelling does not seem the most appropriate at first sight, given the shape of the data histogram. In order to further illustrate the applicability of the skew-Laplace model to various datasets, Figure 3.13 shows the predictive distribution obtained with the skew-Laplace model for three simulated samples. We have drawn $n = 100$ observations from the skew-normal distribution of Azzalini (1985) (panel (a)) and a symmetric Student-t with 2 degrees of freedom (panel (c)). The data in panel (b) were generated from a Gamma(2,5) distribution ($n = 1000$). The skew-normal and gamma data were analysed with a skew-Laplace truncated at zero. For the Student-t sample, we used an untruncated skew-Laplace model. In all cases we have recorded data up to one decimal place (so that $d = 0.05$). Clearly, the skew-Laplace fits the rather different shapes of these three data sets quite well.

Chapter 4

Inference in Two-Piece Location-Scale Models With Jeffreys Priors

"Tyger! Tyger! burning bright
In the forests of the night,
What immortal hand or eye
Dare frame thy fearful symmetry?"
William Blake,
The Tyger.

This chapter addresses the use of Jeffreys priors in the context of univariate three-parameter location-scale models, where skewness is introduced by differing scale parameters either side of the location. We focus on various commonly used parameterisations for these models. In particular, we show the model studied in Chapter 3 can be obtained under a certain reparameterisation of this sort of models. Jeffreys priors are shown to lead to improper posteriors in the wide and practically relevant class of distributions obtained by skewing scale mixtures of normals. Easily checked conditions under which independence Jeffreys priors can be used for valid inference are derived. We also investigate two alternative priors, one of which is shown to lead to valid Bayesian inference for all practically interesting parameterisations of these models and is our recommendation to practitioners. We illustrate some of these models using real data.

4.1 Introduction

The use of skewed distributions is an attractive option for modeling data presenting departures from symmetry. Several mechanisms to obtain skewed distributions by appropriately modifying symmetric distributions have been presented in the literature (Azzalini, 1985; Fernández and Steel, 1998a; Mudholkar and Hutson, 2000).

We focus on the simple univariate location-scale model where we induce skewness by the use of different scales on either sides of the mode and only distinguish three scalar parameters. We investigate Bayesian inference using Jeffreys priors in this simple setting. Despite the simplicity of these models they often fit observed data quite well, and have been used recently in a wide variety of applied contexts, such as genetics, biology, hydrology, economics, finance, medicine, agriculture and marketing (Purdom and Holmes, 2005; Trindade et al., 2010; Rubio and Steel, 2011a; Punathumparambath et al., 2012). For example, they are used for the widely discussed probability forecasts of gross domestic product and inflation produced by the Bank of England and the Sveriges Riksbank (Wallis, 2004; Galbraith and van Norden, 2012). The availability of a "benchmark" Bayesian analysis is thus of particular importance for practitioners.

Firstly, we consider univariate (continuous) two-piece distributions with different scales on both sides of the location parameter. Then, we focus on the family of reparameterisations defined in Arellano-Valle et al. (2005), where the scales are reparameterised in terms of a common scale and a skewness parameter. Whereas we discuss orthogonality of parameterisations, which is of direct interest for likelihood-based frequentist inference, we will mostly focus on Bayesian inference in this chapter. A commonly used prior structure to reflect an absence of prior information is the Jeffreys (or "Jeffreys-rule") prior, which is the reference prior (Berger et al., 2009) in the case of a scalar parameter under asymptotic posterior normality. Under these conditions, Clarke and Barron (1994) showed that this prior asymptotically maximises the expected information from repeated sampling. The Jeffreys prior is an interesting choice because no subjective parameters have to be elicited and it is invariant under reparameterisations (Jeffreys, 1941; Ibrahim and Laud, 1991).

However, in our two-piece location-scale framework (and its reparameterisations), we show that Jeffreys prior does not lead to a proper posterior in the wide and empirically interesting class of distributions obtained by skewing scale mixtures of normals. In addition, we consider the independence Jeffreys prior (constructed as the product of the Jeffreys priors for each parameter while considering the other parameters are fixed), which is shown to lead to a proper posterior under some parameterisations. Simple conditions regarding posterior existence with the independence Jeffreys prior are derived. We propose an alternative prior structure, which is partly subjective, but which is easily elicited and leads to valid Bayesian inference in a wide and practically relevant class of parameterisations of two-piece models.

The structure of this document is as follows: in Section 4.2 we present the two-piece location-scale model and the family of parameterisations defined in Arellano-Valle et al. (2005). We derive the Fisher information matrix for these models as well as the Jeffreys and independence Jeffreys priors. In Section 4.3 we examine posterior existence with these priors in the context of a scale mixture of normals for the underlying symmetric distribution. We also propose two alternative prior structures, one of which is our recommended prior choice for users of these models. In Section 4.4 we present an application of the Bayesian models studied here on a real data set. The final section contains concluding remarks. Proofs of all theorems as well as a numerical coverage analysis of the 95% credible intervals for various models are given in the supplementary material.

4.2 Sampling Models and Jeffreys Priors

4.2.1 Two-piece Location-Scale Models

Let $f(y;\mu,\sigma)$ be an absolutely continuous density with support on $\mathbb{R}$, location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma \in \mathbb{R}^+$, and denote $f\left(\frac{y-\mu}{\sigma};0,1\right) = f\left(\frac{y-\mu}{\sigma}\right)$. Consider the following "two-piece" density constructed from $f\left(\frac{y-\mu}{\sigma_1}\right)$ truncated to $(-\infty,\mu)$ and $f\left(\frac{y-\mu}{\sigma_2}\right)$ truncated to $[\mu,\infty)$:

$$g(y;\mu,\sigma_1,\sigma_2,\varepsilon) = \frac{2\varepsilon}{\sigma_1} f\left(\frac{y-\mu}{\sigma_1}\right) I_{(-\infty,\mu)}(y) + \frac{2(1-\varepsilon)}{\sigma_2} f\left(\frac{y-\mu}{\sigma_2}\right) I_{[\mu,\infty)}(y), \quad (4.1)$$

where $\sigma_1 \in \mathbb{R}^+$ and $\sigma_2 \in \mathbb{R}^+$ are separate scale parameters and $0 < \varepsilon < 1$. In order to get a continuous density, we need to consider the special case where $\varepsilon = \sigma_1/(\sigma_1+\sigma_2)$, so that

$$s(y;\mu,\sigma_1,\sigma_2) = \frac{2}{\sigma_1+\sigma_2}\left[f\left(\frac{y-\mu}{\sigma_1}\right) I_{(-\infty,\mu)}(y) + f\left(\frac{y-\mu}{\sigma_2}\right) I_{[\mu,\infty)}(y)\right]. \quad (4.2)$$

Typically, f will be a symmetric density function. In this chapter, we will assume f to be symmetric with a single mode at zero, which means that μ is the mode of the density in (4.2). If we choose f to be normal and Student densities, the distribution in (4.2) corresponds to split-normal and split-t distributions, respectively, as defined in Geweke (1989). In earlier work, the case with normal f was termed joined half-Gaussian by Gibbons and Mylroie (1973) and two-piece normal by John (1982). A historical account of the many guises of this distribution is provided in Wallis (2013). In line with most of the recent literature (Jones, 2006; Jones and Anaya-Izquierdo, 2010; Wallis, 2013), we shall denote the model in (4.2) as the two-piece model. Since

$$\int_{-\infty}^{\mu} s(y;\mu,\sigma_1,\sigma_2)\,dy = \frac{\sigma_1}{\sigma_1+\sigma_2}, \tag{4.3}$$

s is skewed about μ if $\sigma_1 \neq \sigma_2$ and the ratio σ_1/σ_2 controls the allocation of mass to each side of μ.

We are mainly interested in the inferential properties of these skewed distributions under the popular Jeffreys priors, but will also briefly discuss orthogonality of their parameters. We use the concept of orthogonality in Cox and Reid (1987), which relates to zeroes in the Fisher information matrix of the model. If θ_1 is orthogonal to θ_2, we will denote this as $\theta_1 \perp \theta_2$.

We first calculate the Fisher information matrix and characterise, in terms of the symmetric density f, the cases where this matrix is well defined:

Theorem 7 *Let $s(y;\mu,\sigma_1,\sigma_2)$ be as in (4.2) and suppose that the following conditions hold*

(i) $\int_0^\infty \left[\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt < \infty$,

(ii) $\int_0^\infty t^2 \left[\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt < \infty$,

(iii) $\lim_{t\to\infty} tf(t) = 0$ *or* $\int_0^\infty tf'(t)dt = -\frac{1}{2}$, *which means that* $f(t)$ *is* $o\left(\frac{1}{t}\right)$.

Then the Fisher information matrix $I(\mu,\sigma_1,\sigma_2)$ *is*

$$\begin{pmatrix} \frac{2\alpha_1}{\sigma_1\sigma_2} & -\frac{2\alpha_3}{\sigma_1(\sigma_1+\sigma_2)} & \frac{2\alpha_3}{\sigma_2(\sigma_1+\sigma_2)} \\ -\frac{2\alpha_3}{\sigma_1(\sigma_1+\sigma_2)} & \frac{\alpha_2}{\sigma_1(\sigma_1+\sigma_2)} + \frac{\sigma_2}{\sigma_1(\sigma_1+\sigma_2)^2} & -\frac{1}{(\sigma_1+\sigma_2)^2} \\ \frac{2\alpha_3}{\sigma_2(\sigma_1+\sigma_2)} & -\frac{1}{(\sigma_1+\sigma_2)^2} & \frac{\alpha_2}{\sigma_2(\sigma_1+\sigma_2)} + \frac{\sigma_1}{\sigma_2(\sigma_1+\sigma_2)^2} \end{pmatrix}, \tag{4.4}$$

where

$$\begin{aligned} \alpha_1 &= \int_0^\infty \left[\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt, \\ \alpha_2 &= 2\int_0^\infty \left[1 + t\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt = -1 + 2\int_0^\infty t^2 \left[\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt, \\ \alpha_3 &= \int_0^\infty t\left[\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt. \end{aligned}$$

Conditions (i) and (ii) are required for the existence of the expression in (4.4) and are satisfied under regularity conditions (Lehmann and Casella, 1998). Condition (iii) is useful to simplify some expressions and is satisfied by many models of interest. As examples, normal, Student t, logistic, Cauchy, Laplace and exponential power distributions

(Box and Tiao, 1973) all satisfy $(i)-(iii)$. Given that α_1, α_2 and α_3 are positive as long as $f'(t) \neq 0$ everywhere, none of the entries of the Fisher information matrix are zero. Therefore, this is a non-orthogonal parameterisation.

The Jeffreys prior, proposed by Jeffreys (1941), is defined as the square root of the determinant of the Fisher information matrix. In contrast, the independence Jeffreys prior is defined as the product of the Jeffreys priors for each parameter independently, while treating the others parameters as fixed.

Corollary 1 *If the Fisher information matrix in (4.4) is non-singular, then the Jeffreys prior for the parameters in (4.2) is*

$$\pi_J(\mu, \sigma_1, \sigma_2) \quad \propto \quad \frac{1}{\sigma_1\sigma_2(\sigma_1+\sigma_2)}. \tag{4.5}$$

The independence Jeffreys prior is

$$\pi_I(\mu, \sigma_1, \sigma_2) \quad \propto \quad \frac{\sqrt{[\sigma_1+\alpha_2(\sigma_1+\sigma_2)][\sigma_2+\alpha_2(\sigma_1+\sigma_2)]}}{\sqrt{\sigma_1\sigma_2}(\sigma_1+\sigma_2)^2}. \tag{4.6}$$

The Jeffreys prior is defined only in the cases when the Fisher information matrix is non-singular. The determinant of the Fisher information matrix can be factored into two terms, one dependent on the parameters and the other dependent on the constants $(\alpha_1, \alpha_2, \alpha_3)$. The former is always positive. The following result gives conditions on the density f that ensure that the second factor does not vanish and the Fisher information matrix is thus non-singular.

Theorem 8 *If the conditions of Theorem 7 are satisfied and $f'(t) \neq 0$ $a.e.$, then the Fisher information matrix is non-singular.*

In particular, the Fisher information matrix (4.4) is non-singular if f corresponds to a normal, Laplace, exponential power, logistic, Cauchy or Student t distribution. The structure of the independence Jeffreys prior in (4.6) assumes that $\alpha_2 > 0$, which will always be the case (see the proof of Theorem 8 in the Appendix).

4.2.2 Reparameterisations of the Two-Piece Model

In order to link the two-piece model in (4.2) with the family defined in Arellano-Valle et al. (2005), consider the following reparameterisation (one-to-one transformation)

$$\begin{aligned}(\mu,\sigma_1,\sigma_2) &\leftrightarrow (\mu,\sigma,\gamma), \\ \mu &= \mu, \\ \sigma_1 &= \sigma b(\gamma), \\ \sigma_2 &= \sigma a(\gamma),\end{aligned} \tag{4.7}$$

where $\gamma \in \Gamma$, $\sigma > 0$ and $a(\gamma) > 0$ and $b(\gamma) > 0$ are differentiable functions such that

$$0 < |\lambda(\gamma)| < \infty, \text{with } \lambda(\gamma) \equiv \frac{d}{d\gamma}\log\left[\frac{a(\gamma)}{b(\gamma)}\right]. \tag{4.8}$$

The condition in (4.8) implies that (4.7) is a non-singular mapping and is thus necessary for it to be a one-to-one transformation. Then we get the following reparameterised density from (4.2)

$$s(y;\mu,\sigma,\gamma) = \frac{2}{\sigma[a(\gamma)+b(\gamma)]}\left[f\left(\frac{y-\mu}{\sigma b(\gamma)}\right)I_{(-\infty,\mu)}(y) + f\left(\frac{y-\mu}{\sigma a(\gamma)}\right)I_{[\mu,\infty)}(y)\right]. \tag{4.9}$$

This expression was presented by Arellano-Valle et al. (2005) as a general class of asymmetric distributions, which includes various skewed distributions presented in the literature. Like Jones (2006), we view (4.9) with a given choice of f not as a class of densities but as a class of reparameterisations of the same density.

Two parameterisations in terms of the functions $\{a(\gamma), b(\gamma)\}$ have been widely studied: the inverse scale factors (ISF) model (Fernández and Steel, 1998a), corresponding to $\{a(\gamma), b(\gamma)\} = \{\gamma, 1/\gamma\}$ for $\gamma \in \mathbb{R}^+$ and the ϵ-skew model (Mudholkar and Hutson, 2000), which chooses $\{a(\gamma), b(\gamma)\} = \{1+\gamma, 1-\gamma\}$ for $\gamma \in (-1, 1)$.

The Fisher information matrix for the reparameterised model in (4.9) is as follows:

Theorem 9 *Let $f(y;\mu,\sigma)$ be as in Theorem 7. Then the Fisher information matrix $I(\mu,\sigma,\gamma)$ for model (4.9) is*

$$\begin{pmatrix} \frac{2\alpha_1}{a(\gamma)b(\gamma)\sigma^2} & 0 & \frac{2\alpha_3}{\sigma[a(\gamma)+b(\gamma)]}\left[\frac{a'(\gamma)}{a(\gamma)} - \frac{b'(\gamma)}{b(\gamma)}\right] \\ 0 & \frac{\alpha_2}{\sigma^2} & \frac{\alpha_2}{\sigma}\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right] \\ \frac{2\alpha_3}{\sigma[a(\gamma)+b(\gamma)]}\left[\frac{a'(\gamma)}{a(\gamma)} - \frac{b'(\gamma)}{b(\gamma)}\right] & \frac{\alpha_2}{\sigma}\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right] & \frac{\alpha_2+1}{a(\gamma)+b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)} + \frac{a'(\gamma)^2}{a(\gamma)}\right] - \left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right]^2 \end{pmatrix}.$$

The fact that the elements I_{12} and I_{21} are zero indicates that this reparameterisation is interesting because it induces orthogonality between the parameters μ and σ for any choice of $\{a(\gamma), b(\gamma)\}$. In addition, by appropriately choosing the pair of functions $\{a(\gamma), b(\gamma)\}$ we can generate more zero entries in the Fisher information matrix, as shown in the following corollary.

Corollary 2 *If* $\frac{d}{d\gamma} \log\left[a(\gamma) + b(\gamma)\right] = 0$, *then* $I_{23} = I_{32} = 0$. *In particular if* $a(\gamma) + b(\gamma)$ *is constant, then* $I_{23} = I_{32} = 0$.

If $\alpha_3 > 0$, then $I_{13} = I_{31} = 0$ only if $a(\gamma) \propto b(\gamma)$ which does not satisfy (4.8). Jones and Anaya-Izquierdo (2010) analysed the zeroes of the expectation of the Hessian matrix of (μ, σ, γ) in model (4.9) augmented with an extra parameter to model the properties of f. They also found that $\mu \perp \sigma$ and if $a(\gamma) + b(\gamma)$ is constant then $\sigma \perp \gamma$ as in Corollary 2.

The corresponding Jeffreys prior and independence Jeffreys prior for the parameterisation (μ, σ, γ) are given in the following result.

Corollary 3 *If the Fisher information matrix is non-singular, then the Jeffreys prior for the parameters in (4.9) is*

$$\pi_J(\mu, \sigma, \gamma) \propto \frac{|a'(\gamma)b(\gamma) - a(\gamma)b'(\gamma)|}{\sigma^2 a(\gamma)b(\gamma)[a(\gamma) + b(\gamma)]} = \frac{|\lambda(\gamma)|}{\sigma^2[a(\gamma) + b(\gamma)]}, \tag{4.10}$$

where $\lambda(\gamma)$ *was defined in (4.8). The independence Jeffreys prior is*

$$\pi_I(\mu, \sigma, \gamma) \quad \propto \quad \frac{1}{\sigma}\sqrt{\frac{\alpha_2 + 1}{a(\gamma) + b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)} + \frac{a'(\gamma)^2}{a(\gamma)}\right] - \left[\frac{a'(\gamma) + b'(\gamma)}{a(\gamma) + b(\gamma)}\right]^2}. \tag{4.11}$$

Conditions to ensure non-singularity of the Fisher information matrix for the parameterisation in (4.9) are similar to those obtained for the two-piece model (4.2) in Theorem 8. The only difference is that in this case we have to choose a pair of functions $\{a(\gamma), b(\gamma)\}$ such that (4.7) corresponds to a non-singular transformation:

Corollary 4 *If the conditions of Theorem* 7 *are satisfied,* $f'(t) \neq 0$ *a.e., and (4.8) holds, then the Fisher information matrix corresponding to model (4.9) is non-singular.*

Due to the invariance property of the Jeffreys prior there is a one-to-one relationship between (4.5) and (4.10). On the other hand, the independence Jeffreys prior is not invariant under reparameterisations, so the properties of this prior are dependent on the choice of $\{a(\gamma), b(\gamma)\}$.

Now we will briefly discuss the inverse scale factors and ϵ-skew models.

Inverse Scale Factors Model

The ISF model corresponds to choosing $\{a(\gamma) = \gamma, b(\gamma) = 1/\gamma\}$, $\gamma \in \mathbb{R}^+$ in (4.9), so that from Theorem 9 the Fisher information matrix of the parameters (μ, σ, γ) is

$$I(\mu,\sigma,\gamma) = \begin{pmatrix} \frac{2\alpha_1}{\sigma^2} & 0 & \frac{4\alpha_3}{\sigma(\gamma^2+1)} \\ 0 & \frac{\alpha_2}{\sigma^2} & \frac{\alpha_2\left(\gamma^2-1\right)}{\sigma(\gamma^3+\gamma)} \\ \frac{4\alpha_3}{\sigma(\gamma^2+1)} & \frac{\alpha_2\left(\gamma^2-1\right)}{\sigma(\gamma^3+\gamma)} & \frac{\alpha_2}{\gamma^2} + \frac{4}{(\gamma^2+1)^2} \end{pmatrix}. \tag{4.12}$$

If the Fisher information matrix in (4.12) is non-singular, then the Jeffreys prior for the ISF model is

$$\pi_J(\mu,\sigma,\gamma) \quad \propto \quad \frac{1}{\sigma^2\,(1+\gamma^2)}, \tag{4.13}$$

which has a finite integral over $\gamma \in \mathbb{R}^+$, but is improper in terms of μ and σ. The independence Jeffreys prior is

$$\pi_I(\mu,\sigma,\gamma) \quad \propto \quad \frac{1}{\sigma}\sqrt{\frac{\alpha_2}{\gamma^2} + \frac{4}{(\gamma^2+1)^2}}, \tag{4.14}$$

which is not integrable in any of the parameters.

ϵ-Skew Model

For the ϵ-skew model we choose $\{a(\gamma) = 1-\gamma, b(\gamma) = 1+\gamma\}$ in (4.9), where $\gamma \in (-1, 1)$, leading to the Fisher information matrix

$$I(\mu,\sigma,\gamma) = \begin{pmatrix} \frac{2\alpha_1}{\sigma^2(1-\gamma^2)} & 0 & -\frac{2\alpha_3}{\sigma(1-\gamma^2)} \\ 0 & \frac{\alpha_2}{\sigma^2} & 0 \\ -\frac{2\alpha_3}{\sigma(1-\gamma^2)} & 0 & \frac{\alpha_2+1}{1-\gamma^2} \end{pmatrix}. \tag{4.15}$$

The ϵ-skew parameterisation satisfies the condition in Corollary 2 and thus its Fisher information matrix has four zeroes. This feature simplifies classical inference. For example, in the cases where f is normal or Laplace, the corresponding ϵ-skew model leads to maximum likelihood estimators in closed form (Mudholkar and Hutson, 2000; Arellano-Valle et al., 2005).

Provided the Fisher information matrix in (4.15) is non-singular, the Jeffreys prior for the ϵ-skew model is

$$\pi_J(\mu, \sigma, \gamma) \quad \propto \quad \frac{1}{\sigma^2(1-\gamma^2)}, \tag{4.16}$$

which is not integrable in any of the parameters. The independence Jeffreys prior is

$$\pi_I(\mu, \sigma, \gamma) \quad \propto \quad \frac{1}{\sigma\sqrt{1-\gamma^2}}, \tag{4.17}$$

which has a finite integral over $\gamma \in (-1, 1)$, but does not integrate in μ and σ. For this model the independence Jeffreys prior does not depend on f (through α_2), in contrast with the priors for the two-piece model in (4.6) and the ISF model in (4.14).

In the different models mentioned above, the skewness parameter γ does not have the same interpretation. This makes it particularly difficult to compare models and priors on γ. It is therefore helpful to introduce a measure of skewness which has a common meaning for all models. In particular, we use the skewness measure with respect to the mode from Arnold and Groeneveld (1995), defined as

Definition 5 *The Arnold-Groeneveld measure of skewness for a distribution function S corresponding to a unimodal density with the mode at M is defined as 1 minus twice the probability mass to the left of the mode:*

$$AG = 1 - 2S(M).$$

The AG measure takes values in $(-1, 1)$ and provides information about the allocation of mass to each side of the mode. Positive values of AG indicate right skewness while negative values indicate left skewness. From (4.3) it is immediate that for the two-piece model $AG = (\sigma_2 - \sigma_1)/(\sigma_1 + \sigma_2)$, which only depends on the two scales and not on the properties of f. Similarly, for the parameterisation in Arellano-Valle et al. (2005) in (4.9) the AG skewness measure has a closed form which only depends on γ:

$$AG(\gamma) = \frac{a(\gamma) - b(\gamma)}{a(\gamma) + b(\gamma)}.$$

For the special case of the ISF model in Subsection 4.2.2, this reduces to

$$AG(\gamma) = \frac{\gamma^2 - 1}{\gamma^2 + 1},$$

while for the ϵ-skew model in Subsection 4.2.2 we obtain $AG(\gamma) = -\gamma$.

In both examples above, the AG skewness measure is a monotonic function of γ, so we can meaningfully interpret γ as a skewness parameter. In general, we will be mostly interested in parameterisations such that this is the case, which can be characterised as follows:

Theorem 10 *Let s, $a(\gamma)$ and $b(\gamma)$ be as in (4.9), then for any unimodal density f*

- *$AG(\gamma)$ is increasing if and only if $\lambda(\gamma) > 0$.*
- *$AG(\gamma)$ is decreasing if and only if $\lambda(\gamma) < 0$.*

4.3 Inference

In this section we will present necessary and/or sufficient conditions for the properness of the posterior distribution of the parameters of the two-piece models considered when using the priors presented in the previous section, as well as two alternative priors to be introduced later in Subsection 3.4. Throughout this section we will assume that we have observed a sample of n independent replications from either (4.2) or (4.9). We separately deal with samples where all the observations are different and samples which contain repeated observations. Most of the results in this section are for the case where the underlying symmetric distribution (with density f) belongs to the wide class of scale mixtures of normals. Of course, a meaningful use of the results in Subsections 3.1 and 3.2 implies a nonsingular information matrix (see Theorem 8 and Corollary 4) so that the Jeffreys prior exists or a well-defined independence Jeffreys prior. However, most cases of practical interest will correspond to an f that allows for these priors to be well-defined.

Recall that a density f corresponds to a scale mixture of normals if it can be written as

$$f(x;\nu) = \int_0^\infty \tau^{1/2}\phi(\tau^{1/2}x)dP_{\tau|\nu},$$

where ϕ is the standard normal density and $P_{\tau|\nu}$ is a mixing distribution on $\mathbb{R}_+$. The class of scale mixtures of normals is quite a rich class of symmetric and unimodal continuous distributions and contains many popular distributions, such as the normal, Student t with ν degrees of freedom, logistic, Laplace, Cauchy and the exponential power family with power $1 \leq q < 2$ (see Fernández and Steel, 2000 for more details). This class does not cover distributions with tails thinner than normal tails.

4.3.1 Independence Jeffreys Prior

The independence Jeffreys prior is not invariant under reparameterisations. Therefore if we consider one-to-one transformations as in (4.7), we need to analyse the properness of the posterior distribution of (μ, σ, γ) for each specific choice of $\{a(\gamma), b(\gamma)\}$. Thus, we consider the models in (4.2) and (4.9) separately.

Theorem 11 *Let* $\mathbf{y} = (y_1, \ldots, y_n)$ *be an independent sample from the model in* (4.2)*, where* f *is a scale mixture of normals. Then,*

(i) The posterior distribution of $(\mu, \sigma_1, \sigma_2)$ *using the independence Jeffreys prior (4.6) is proper if* $n \geq 2$ *and all the observations are different.*

(ii) Suppose that the sample $\mathbf{y}$ *contains repeated observations. Let* k *be the largest number of observations with the same value in* $\mathbf{y}$*. If* $1 < k < n$*, then the posterior of* $(\mu, \sigma_1, \sigma_2)$ *is proper if and only if the mixing distribution of* f *satisfies*

$$\int_{0<\tau_1 \leq \cdots \leq \tau_n < \infty} \tau_{n-k}^{-(n-2)/2} \prod_{i \neq n-k,n} \tau_i^{1/2} dP_{(\tau_1,\ldots,\tau_n)} < \infty. \tag{4.18}$$

In the case of the two-piece normal sampling model (i.e. normal f*), it suffices to have two different observations.*

Thus, for a wide and practically important class of distributions f, the two-piece model in (4.2) with the independence Jeffreys prior leads to valid inference in any sample of two or more observations.

For the model in (4.9), we can derive useful existence results within a class of prior distributions:

Theorem 12 *Let* $\mathbf{y} = (y_1, \ldots, y_n)$ *be an independent sample from the model in* (4.9)*, where* f *is a scale mixture of normals. Consider a prior distribution of the form* $\pi(\mu, \sigma, \gamma) \propto \sigma^{-1}\pi(\gamma)$*, for some* $\pi(\gamma)$*. Then:*

(i) a necessary condition for the properness of the posterior distribution of (μ, σ, γ) *is*

$$\int_{\Gamma} \left[\frac{a(\gamma)}{a(\gamma) + b(\gamma)}\right]^n \pi(\gamma)\, d\gamma < \infty. \tag{4.19}$$

(ii) the posterior distribution of (μ, σ, γ) *is proper if* $n \geq 2$*, all the observations are different, and* $\pi(\gamma)$ *is proper.*

(iii) Suppose that the sample $\mathbf{y}$ *contains repeated observations and* $\pi(\gamma)$ *is proper. Let* k *be the largest number of observations with the same value in* $\mathbf{y}$*. If* $1 < k < n$*, then the posterior of* (μ, σ, γ) *is proper if and only if the mixing distribution of* f *satisfies* (4.18)*. In the case of the two-piece normal sampling model (i.e. normal* f*), it suffices to have two different observations.*

This theorem implies that a posterior will exist for the ϵ-skew model under the independence Jeffreys prior in (4.17), as this prior is a member of the class in Theorem 12 with proper $\pi(\gamma)$.

However, for the ISF model the independence Jeffreys prior does not integrate in γ and we can show that the necessary condition (4.19) is violated, so that a posterior does not exist in this case:

Corollary 5 *If f is a scale mixture of normals in (4.9) and $\{a(\gamma), b(\gamma)\}$ are as in the inverse scale factors model, then the posterior distribution of (μ, σ, γ) is improper under the independence Jeffreys prior (4.14).*

Theorem 12 emphasises the relevance of the choice of the functions $\{a(\gamma), b(\gamma)\}$ for the properness of the posterior distribution of (μ, σ, γ) when using the independence Jeffreys prior. In particular, condition (4.19) can be used to detect parameterisations $\{a(\gamma), b(\gamma)\}$ that produce improper posteriors. The fact that the ISF model does not allow for inference with the independence Jeffreys prior is rather surprising since this prior almost always leads to proper posteriors, and the ISF model is quite a straightforward extension of the usual location-scale model. Subsection 4.3.3 will shed more light on this.

4.3.2 Jeffreys Prior

We now examine the properness of the posterior distribution of the parameters (μ, σ, γ) under the Jeffreys prior. An important feature of this prior is the invariance under one-to-one reparameterisations. Therefore, the results regarding the properness of the posterior of (μ, σ, γ) for any choice of $\{a(\gamma), b(\gamma)\}$ in model (4.9) that corresponds to a one-to-one transformation in (4.7) are the same and also applicable to the posterior of $(\mu, \sigma_1, \sigma_2)$ in model (4.2).

Theorem 13 *Let s be as in (4.9), assume that f is a scale mixture of normals and consider the Jeffreys prior (4.10) for the parameters of this model. Then, for $n \geq 2$, a necessary condition for the properness of the posterior distribution of (μ, σ, γ) is*

$$\int_{\Gamma} \left[\frac{a(\gamma)}{a(\gamma) + b(\gamma)}\right]^{n+1} |\lambda(\gamma)| \, d\gamma < \infty, \tag{4.20}$$

with $\lambda(\gamma)$ defined as in (4.8).

Corollary 6 *Consider sampling from (4.9) with f a scale mixture of normals and $\{a(\gamma), b(\gamma)\}$ as in the inverse scale factors model, then the posterior distribution of (μ, σ, γ) is improper using the Jeffreys prior (4.10). As a consequence, for any pair of functions $\{a(\gamma), b(\gamma)\}$*

such that the mapping $(\mu, \sigma_1, \sigma_2) \leftrightarrow (\mu, \sigma, \gamma)$ is one-to-one, the posterior distribution of (μ, σ, γ) is improper using the Jeffreys prior (4.10).
Proof. *We can verify that the necessary condition (4.20) is not satisfied for these functions.*

This corollary implies that we can not conduct Bayesian inference for the parameters of this type of skewed distributions using the Jeffreys prior. It is rather rare to find that the Jeffreys prior does not lead to a proper posterior, and it is somewhat surprising to find that we can not use this prior in these rather simple classes of two-piece distributions with only three parameters.

Because the Jeffreys prior is invariant to reparameterisation, its use is thus prohibited in any one-to-one reparameterisation of the two-piece models in (4.2) or (4.9). However, one way to get around this problem is to choose functions $\{a(\gamma), b(\gamma)\}$ such that the mapping $(\mu, \sigma, \gamma) \mapsto (\mu, \sigma_1, \sigma_2)$ is not one-to-one, but hopefully still of some interest for modelling. Another way to produce a proper posterior distribution when using the Jeffreys prior is to restrict Γ such that $\lambda(\gamma)$ is absolutely integrable.

Theorem 14 *Let s be as in (4.9) where f is normal or Laplace. Consider the Jeffreys prior (4.10) for the parameters of this model. Let $\{a(\gamma), b(\gamma)\}$ be continuously differentiable functions for $\gamma \in \Gamma$ such that*

$$0 < \int_\Gamma |\lambda(\gamma)|\, d\gamma < \infty. \tag{4.21}$$

Then we have the following results

(i) The posterior distribution of (μ, σ, γ) is proper when $n \geq 2$ and there are at least two different observations.

(ii) The mapping $(\mu, \sigma, \gamma) \mapsto (\mu, \sigma_1, \sigma_2)$ is not one-to-one.

(iii) If Γ is an interval (not necessarily bounded) and $AG(\gamma)$ is monotonic, then $AG(\gamma)$ is not surjective.

First, we considered forcing existence of the posterior through the choice of the functions $\{a(\gamma), b(\gamma)\}$, in particular such that the ratio $a(\gamma)/b(\gamma)$ is bounded, which excludes a one-to-one reparameterisation in (4.7). However, the examples we generated in this way did not lead to implied priors on AG that could be of interest to practitioners.

It is actually easier to generate examples of practical relevance if we consider restricting the parameter space of γ in the context of functions $\{a(\gamma), b(\gamma)\}$ that would not lead to a proper posterior with unrestricted γ. The following is such an example.

Example 6 (Logistic AG) *Consider* $a(\gamma) = 1+\exp(2\gamma)$, $b(\gamma) = 1+\exp(-2\gamma)$ *for* $\gamma \in \mathbb{R}$, *then*

$$\begin{aligned} AG(\gamma) &= \tanh(\gamma), \\ \lambda(\gamma) &= 2 \\ \pi_J(\mu,\sigma,\gamma) &\propto \frac{1}{\sigma^2}\operatorname{sech}(\gamma)^2. \end{aligned} \tag{4.22}$$

In addition, the functions $a(\gamma)$, $b(\gamma)$ *and* $AG(\gamma)$ *are monotonic* $\forall\, \gamma \in \mathbb{R}$, *the Jeffreys prior in (4.22) implies that* $AG \sim \mathrm{Unif}(-1,1)$ *and* $AG : \mathbb{R} \mapsto (-1,1)$. *Clearly,* $\lambda(\gamma)$ *is not integrable on* $\mathbb{R}$, *but if we restrict* $\gamma \in [-B,B]$ *for some* $0 < B < \infty$, *then we can use the Jeffreys prior (4.22) for making inference on* (μ,σ,γ) *for normal or Laplace* f *and* $AG : \mathbb{R} \mapsto [\tanh(-B), \tanh(B)]$. *Figure 4.1 presents the functions* $a(\gamma)$, $b(\gamma)$ *and* $AG(\gamma)$ *for* $B = 3$. *The induced prior on* AG *is a Uniform over the range* $[\tanh(-B), \tanh(B)] = [-0.995, 0.995]$.

We will call the model in Example 6 the "logistic AG model" as AG(γ) is a logistic function of γ transformed to take values in the interval (-1,1) for $\gamma \in \mathbb{R}$. The choice of $a(\gamma)$ and $b(\gamma)$ does lead to a one-to-one transformation in (4.7) when $\gamma \in \mathbb{R}$, but not if γ is restricted to a bounded interval: then the ratio $a(\gamma)/b(\gamma)$ is also bounded and this precludes a one-to-one mapping. $a(\gamma)$ and $b(\gamma)$ satisfy the condition $a(\gamma) + b(\gamma) = a(\gamma)b(\gamma)$, which induces a really interesting structure on the Jeffreys prior, namely that it implies a uniform prior in terms of the AG measure. This might be an attractive prior for practitioners to use in the absence of strong prior information.

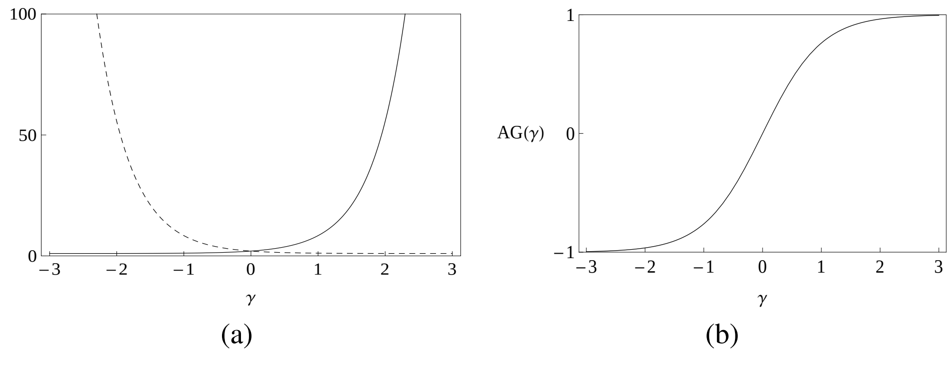


Figure 4.1: (a) $a(\gamma)$ (solid line) and $b(\gamma)$ (dashed line); (b) $AG(\gamma)$.

4.3.3 Intuitive Explanation

As mentioned before, the lack of a posterior under a commonly used prior in what is essentially a very simply generalisation of a standard location-scale model can be considered surprising. Thus, we offer a few explanatory comments in this subsection. These are not meant to be formal proofs (they can be found in the Appendix), but merely intuitive ideas that help us understand what drives the main results we have found in the previous subsections.

In the context of the two-piece model in (4.2), it is easy to see that as σ_1 tends to zero, the sampling density tends to the half density on $[\mu, \infty)$ with scale σ_2. Thus, the likelihood will be constant in σ_1 in the neighbourhood of zero. This means the prior needs to integrate in that neighbourhood for a posterior to exist. If we consider the independent Jeffreys prior in (4.6) it behaves like $\sigma_1^{-1/2}$ for small σ_1 and this integrates close to zero. Indeed, we have a posterior in this case. However, the Jeffreys prior in (4.5) behaves like $1/\sigma_1$ for small σ_1 and this does not integrate, thus precluding a posterior. Of course, similar arguments hold for small σ_2.

In the case of the reparameterised model in (4.9), we have a potential problem if one of the scales, say, $\sigma a(\gamma)$ goes to zero. If then the ratio $b(\gamma)/a(\gamma)$ has an upper bound, this will necessarily imply that both scales tend to zero, so the model behaves like a standard location-scale model which leads to a proper posterior under the Jeffreys prior. This is the case explored in Theorem 14 and Example 1. If, however, the ratio between the functions $a(\gamma)$ and $b(\gamma)$ is not bounded and (4.7) defines a one-to-one mapping, we will have no posterior with the Jeffreys prior due to the invariance of this prior under reparameterisation, and it depends on the particular choice of functions $\{a(\gamma), b(\gamma)\}$ whether the independence Jeffreys prior will lead to a posterior. It is helpful to transform the parameters back to those of the two-piece model in (4.2). Then, for the ϵ-skew model the independence Jeffreys prior in (4.17) can be shown to behave like $\sigma_i^{-1/2}$ for small $\sigma_i, i = 1, 2$, which is integrable close to zero, and the posterior is well-defined. On the other hand, the independence Jeffreys prior for the ISF model in (4.14) behaves like $1/\sigma_i$ for small $\sigma_i, i = 1, 2$, which does not integrate in a neighbourhood of zero and precludes posterior existence.

4.3.4 Alternative Priors

We now consider two alternative priors for the sampling model in (4.9): one is a modification of the Jeffreys prior and the other is a non-objective prior with an elicitation strategy in terms of an easily interpretable quantity and the possibility to use vague priors. Both prior structures will be of the form

$$\pi(\mu, \sigma, \gamma) \propto \sigma^{-1}\pi(\gamma). \tag{4.23}$$

Modified Jeffreys Prior

Combining the structure of the independence Jeffreys prior in terms of σ with that of the Jeffreys prior for γ leads to the following modified Jeffreys prior

$$\begin{aligned} \pi_M(\gamma) &\propto \frac{|a'(\gamma)b(\gamma) - a(\gamma)b'(\gamma)|}{a(\gamma)b(\gamma)[a(\gamma)+b(\gamma)]} \\ &= \frac{1}{a(\gamma)+b(\gamma)}\left|\frac{d}{d\gamma}\log\left[\frac{a(\gamma)}{b(\gamma)}\right]\right|. \end{aligned} \tag{4.24}$$

This prior can also be interpreted as the independence Jeffreys prior with the independence applied to the two blocks μ and (σ, γ), rather than the three parameters separately (see Fonseca et al., 2008 for a similar prior in the context of a Student-t regression model with unknown degrees of freedom).

AG **Beta Prior**

The second alternative prior $\pi_\beta(\gamma)$ is such that $\delta = (AG+1)/2$, the AG skewness measure rescaled to the unit interval, has a Beta(α_0, β_0) distribution. Thus, this prior is not obtained through a formal rule and can be elicited on the basis of AG, which has a clear interpretation in terms of probability mass on both sides of the mode (see Definition 5). In practice, this prior is perhaps most useful for values of α_0 and β_0 relatively close to one, reflecting vague prior information on the AG measure of skewness. In terms of γ, it corresponds to

$$\pi_\beta(\gamma) \propto \frac{|a'(\gamma)b(\gamma) - a(\gamma)b'(\gamma)|}{[a(\gamma)+b(\gamma)]^{\alpha_0+\beta_0}} a(\gamma)^{\alpha_0-1}b(\gamma)^{\beta_0-1}. \tag{4.25}$$

Despite being motivated in rather different ways, both alternative priors coincide in certain special cases. In particular, prior (4.24) implies that $\delta \sim$ Beta$(1/2, 1/2)$ if $a(\gamma)b(\gamma) = c$. This is the case of the Inverse Scale Factors parameterisation. In addition, the prior distributions (4.24) and (4.25) coincide if $\alpha_0 = \beta_0 = 1$ and $a(\gamma) + b(\gamma) = a(\gamma)b(\gamma)$, as already remarked in the context of the logistic AG model in Example 1.

The alternative priors of (μ, σ, γ) for the Inverse Scale Factors model are respectively

$$\pi_M(\mu,\sigma,\gamma) \quad \propto \quad \frac{1}{\sigma\,(1+\gamma^2)}, \tag{4.26}$$

$$\pi_\beta(\mu,\sigma,\gamma) \quad \propto \quad \frac{\gamma^{2\alpha_0-1}}{\sigma\,(1+\gamma^2)^{\alpha_0+\beta_0}}, \tag{4.27}$$

for $\gamma \in \mathbb{R}^+$. Indeed both priors coincide when $\alpha_0 = \beta_0 = 1/2$.

In the case of the ϵ-skew model the alternative priors are

$$\pi_M(\mu,\sigma,\gamma) \quad \propto \quad \frac{1}{\sigma(1-\gamma^2)}, \tag{4.28}$$

$$\pi_\beta(\mu,\sigma,\gamma) \quad \propto \quad \frac{(1-\gamma)^{\beta_0-1}(1+\gamma)^{\alpha_0-1}}{\sigma}, \tag{4.29}$$

for $\gamma \in (-1,1)$. The modified Jeffreys prior does not integrate in γ (like the Jeffreys prior), and only coincides with the AG beta prior in the limit as both α_0 and β_0 tend to zero. This could be argued to be a rather counterintuitive prior on AG, putting lots of mass at the extremes.

The alternative priors for the logistic AG parameterisation of Example 1 are

$$\pi_M(\mu,\sigma,\gamma) \quad \propto \quad \frac{1}{\sigma}\operatorname{sech}(\gamma)^2, \tag{4.30}$$

$$\pi_\beta(\mu,\sigma,\gamma) \quad \propto \quad \frac{1}{\sigma}\frac{\left(1+e^{2\gamma}\right)^{\alpha_0}\left(1+e^{-2\gamma}\right)^{\beta_0}}{[1+\cosh(2\gamma)]^{\alpha_0+\beta_0}}, \tag{4.31}$$

for $\gamma \in \mathbb{R}^+$. As mentioned above, for $\alpha_0 = \beta_0 = 1$ both priors coincide.

Since the modified Jeffreys prior $\pi_M(\cdot)$ is not the Jeffreys prior, the parameterisation matters. Whenever the two alternative priors coincide in the examples above, $\pi_M(\cdot)$ corresponds to a symmetric prior in AG, which could be considered "vague" in a rather intuitive sense except for the ϵ-skew case, where the modified Jeffreys prior implies a rather extreme prior when viewed in terms of AG.

Inference

Since this prior structure is of the form (4.23), Theorem 12 presents necessary and sufficient conditions for the properness of the posterior distribution of (μ,σ,γ).

Corollary 7 *Consider sampling from (4.9) where f is a scale mixture of normals. For the Inverse Scale Factors and the logistic AG models the posterior distribution of (μ,σ,γ) using the modified Jeffreys priors (4.26) and (4.30), respectively, is proper if $n \geq 2$ and all*

the observations are different. If $k > 1$ is the largest number of repeated observations in the sample, we have a proper posterior if the mixing distribution of f also satisfies (4.18).
Proof. *Follows from Theorem 12(ii) and (iii) given that these priors imply a proper $\pi(\gamma)$.*

The following corollary illustrates that when using the modified Jeffreys prior, the choice of the functions $\{a(\gamma), b(\gamma)\}$ is critical.

Corollary 8 *The posterior distribution under the modified Jeffreys prior (4.28) in the sampling model (4.9) with f a scale mixture of normals is improper for the ϵ-skew model.*
Proof. *In this case, the necessary condition (4.19) is not satisfied.*

However, for the AG beta prior all three model specifications considered here lead to proper posteriors. In fact, posterior existence is guaranteed within a large class of parameterisations $\{a(\gamma), b(\gamma)\}$, namely all parameterisations for which γ is a one-to-one transformation of AG.

Theorem 15 *Let $y = \{y_1, ..., y_n\}$ be a sample from (4.9) where f is a scale mixture of normals. Consider the AG beta prior in (4.23) and (4.25) with $\alpha_0, \beta_0 > 0$. Then, for any choice $\{a(\gamma), b(\gamma)\}$ such that $\lambda(\gamma)$ defined in (4.8) does not change sign over $\gamma \in \Gamma$ the posterior distribution of (μ, σ, γ) is proper if $n \geq 2$ and all the observations are different. If $k > 1$ is the largest number of repeated observations in the sample, we have a proper posterior if the mixing distribution of f also satisfies (4.18).*

This result means that for all parameterisations for which γ can be considered a skewness parameter (*i.e.* all choices of $\{a(\gamma), b(\gamma)\}$ of practical modelling interest), we will be able to conduct Bayesian inference with the AG beta prior.

4.4 Example

Consider the problem of estimating $\theta = P(X < Y)$. The case when X and Y follow independent normal or exponential distributions has been recently studied, using Jeffreys priors, by Ventura and Racugno (2011). Now, suppose that X and Y are independent variables from univariate two-piece location-scale models as in (4.9) with parameters $(\mu_x, \sigma_x, \gamma_x)$ and $(\mu_y, \sigma_y, \gamma_y)$ respectively. We use the data presented in Heinz et al. (2003). This data set contains the body mass index (BMI) of 260 women and 247 men, who are physically active with ages ranging in the twenties and early thirties. Figure 4.2 shows the histograms of females and males separately. The shape of the histograms suggests the presence of skewness. Therefore, we model these observations with (4.9), using a normal f.

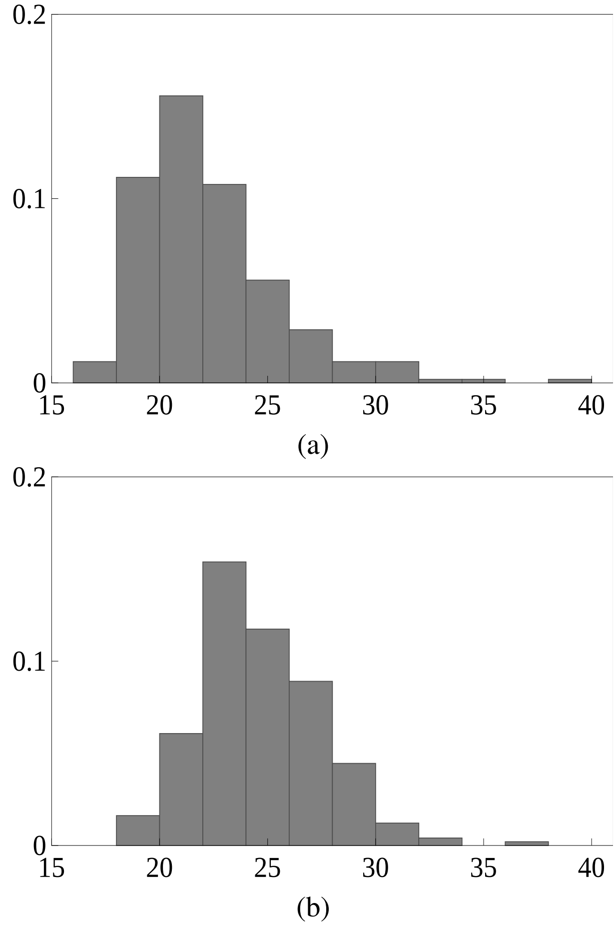


Figure 4.2: Histograms of body mass index data: (a) females; (b) males.

It has been noted that BMI presents a sexual dimorphism and that men tend to have larger BMI than women. Here, we explore this idea through the posterior distribution of θ. We use Models 1 to 5 described in the previous subsection as well as the skew-normal model of Azzalini (1985), which will be denoted as Model 6, given by

$$s(y;\mu,\sigma,\lambda) = \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\lambda\frac{y-\mu}{\sigma}\right),$$

using the prior

$$\pi(\mu,\sigma,\lambda) \propto \sigma^{-1}\pi(\lambda). \tag{4.32}$$

The structure of this prior, using the Jeffreys prior of λ derived in the model without location and scale parameters for $\pi(\lambda)$, was proposed in Liseo and Loperfido (2006), who also prove existence of the posterior under this prior. Bayes and Branco (2007) show that the Jeffreys prior of λ can be approximated by a Student t distribution with $1/2$ degrees of freedom, which is what was used for our calculations.

Using a Markov chain Monte Carlo algorithm, a sample of size $10,000$ was recorded from the posterior distribution after a burn-in period of $50,000$ draws with a thinning of 100 draws for all models. Figure 4.3 presents the posterior distributions of θ.

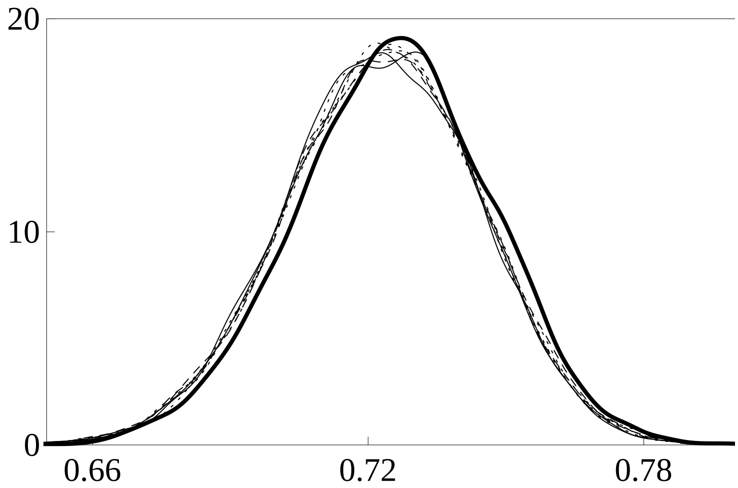


Figure 4.3: Posterior distributions of θ: Models 1 and 2 (continuous lines); Model 3 with $B = 3$, $B = 10$ and $B = 30$ (dotted lines); Models 4 and 5 (dashed lines); Model 6 (bold line).

Clearly, inference with all these different models is very similar, with only the Azzalini model (Model 6) leading to slightly different results. None of the 95% posterior credible intervals include the value $\theta = 0.5$ (in fact the 2.5th percentile is 0.68 for all models), which is in line with the idea that men tend to have larger BMI than women.

4.5 Concluding Remarks

We consider the class of univariate continuous two-piece distributions, which are often used as modifications of the symmetric location-scale model to allow for skewness, and its reparameterised versions as presented in Arellano-Valle et al. (2005), where we can identify a location, a scale and a skewness parameter. A number of well-known models (the inverse scale factor or ISF model and the ϵ-skew model) correspond to particular choices of this parameterisation. In particular, we focus on Bayesian inference in these models using Jeffreys or the independence Jeffreys prior. We prove that these models do not lead to valid posterior inference under Jeffreys prior for any underlying symmetric distribution in the class of scale mixture of normals. As an ad-hoc fix, we show that modifying Jeffreys prior by truncating the support of the skewness parameter can lead to posterior existence. A more fundamental solution is to use the independence Jeffreys prior instead, which is shown to lead to a valid posterior for some parameterisations of these sampling models. However, this is not the case for the ISF model. Two alternative priors are proposed. A modified Jeffreys prior does lead to a posterior for the ISF model, but not for the ϵ-skew model. A second alternative prior is induced by a Beta prior on the AG skewness measure, and is shown to lead to valid inference in a wide class of parameterisations of these models, including the ISF and ϵ-skew models and arguably all models of practical importance. We apply the models, as well as an alternative skewed distribution due to Azzalini (1985), to some real data. For a number of models that lead to valid inference, we compute empirical coverage probabilities of the posterior credible intervals (see the Appendix). This indicates a mostly satisfactory behaviour.

It is important to stress that the three-parameter sampling models examined here are quite simple modifications of the standard location-scale model, and that the Jeffreys prior is a very commonly used prior in the absence of subjective prior information. The fact that the combination of these sampling models with a Jeffreys prior does not lead to a proper posterior is somewhat surprising and definitely relevant for statistical practice, as these models seem attractive options to deal with skewed data, and are used frequently in a wide variety of applied contexts. The better properties of the independence Jeffreys prior are in line with statistical folklore: Jeffreys (1961, p. 182) himself preferred this prior for location-scale problems, and in the univariate normal case the independence Jeffreys is a matching prior (Berger and Sun, 2008). Even with this prior, however, problems of posterior existence can occur, depending on which parameterisation we choose. Two alternative priors are examined, and we recommend the AG beta prior for use with two-piece distributions as it ensures posterior inference for any parameterisation of practical interest and avoids inducing extreme prior beliefs on the easily interpreted AG skewness measure. Using this prior structure we can induce vague or flat priors on the AG measure of skewness, which is a key function of interest of the model parameters in this context (see Seaman III et al., 2012 for a more general discussion of this principle). The AG beta prior is not an objectively obtained prior (even though it has such an interpretation in special cases), but is easily elicited in practice on the basis of a readily interpretable skewness measure.

Chapter 5

Bayesian Inference for $\mathbb{P}(X < Y)$ Using Asymmetric Dependent Distributions

> "Symmetry, although mathematically fascinating, also has a coldness, a rigidity, a fixity, a sense of stasis, which is less interesting, less attractive, indeed less beautiful than asymmetry".
> I. C. McManus,
> *Symmetry and asymmetry in aesthetics and the arts*.

This chapter studies Bayesian inference for $\theta = \mathbb{P}(X < Y)$ in the case where the marginal distributions of X and Y belong to classes of distributions obtained by skewing scale mixtures of normals. We separately address the cases where X and Y are independent or dependent random variables. Dependencies between X and Y are modelled using a Gaussian copula. Noninformative benchmark and vague priors are provided for these scenarios and conditions for the existence of the posterior distribution of θ are presented. We show that the use of the Bayesian models proposed here is also valid in the presence of set observations. Examples using simulated and real data sets are presented.

5.1 Introduction

Stress–strength models have attracted the attention of statisticians for many years due to their applicability in diverse areas such as medicine, engineering, quality control, among others. For example, if X and Y are the outcomes of a treatment and a control group, respectively, then the quantity $\theta = \mathbb{P}(X < Y)$ can be interpreted as the effectiveness of the treatment (Kotz et al., 2003; Ventura and Racugno, 2011). Another important use of $\theta = \mathbb{P}(X < Y)$ in medicine is related to the analysis of receiver operating characteristic (ROC) curves, where θ naturally appears as an index of diagnostic accuracy (Zhou, 2008). The parameter θ can be seen as a function of the parameters of the distribution of the random vector (X, Y) and can be calculated in closed form for a limited number of cases (Kotz et al., 2003; Nadarajah, 2005; Genç, 2012). There is a large amount of literature about the estimation of θ using different approaches and distributional assumptions on (X, Y) (e.g. Kotz et al., 2003, Greco and Ventura, 2011 and Ventura and Racugno, 2011). For instance, it has been assumed that

(i) X and Y are independent (Zhou, 2008; Ventura and Racugno, 2011).

(ii) The distributions of X and Y share common parameters (Gupta and Peng, 2009).

(iii) The distributions of X and Y are independent skewed normals (Azzalini and Chiogna, 2004; Gupta and Brown, 2001).

(iv) X and Y are dependent with a bivariate normal distribution (Nandi and Aich, 1994; Barbiero, 2012)

(v) X and Y are conditionally (on certain unobservable variables) independent exponential random variables (Shoukri et al., 2005).

Although closed expressions for the profile likelihood and modified profile likelihood of θ have been calculated for some particular cases (Montoya, 2008; Ventura and Racugno, 2011; Díaz-Francés and Montoya, 2012), it is difficult (if at all feasible) in the general case to find a reparameterisation of the model parameters that involves θ (Azzalini and Chiogna, 2004; Díaz-Francés and Montoya, 2012). This complicates the calculation of the profile likelihood of the parameter θ, and therefore, the interval estimation using the classical approach.

Alternative inferential approaches for estimating this parameter have also been proposed; for example, the use of confidence intervals (see Kotz et al., 2003), asymptotic confidence intervals and bootstrap (Zhou, 2008), Bayesian inference using reference priors (Sun et al., 1998), nonparametric estimators using kernel methods (Baklizi and Eidous, 2006), and Jackknife empirical likelihoods (Jing et al., 2009). Ventura and Racugno (2011) consider modified profile likelihoods and Bayesian inference using matching priors (see *e.g.* Datta and Ghosh, 1995 for a more general discussion on matching priors). Most of these approaches were proposed under specific distributional assumptions.

To our knowledge, there is a gap in the cases analysed in the literature. The case where X and Y are dependent and the case where their marginal distributions are skewed with support on $\mathbb{R}$ have been analysed separately. This chapter tries to fill this gap by analysing the case where X and Y are dependent with marginal distributions belonging to the class of distributions obtained by skewing scale mixtures of normals. In addition, we address this problem in the context of set observations, which can immediately account for censoring.

In Section 5.2, we study the case where X and Y are independent with particular focus on the case where their distributions are skewed. We consider skewed distributions obtained with two different skewing mechanisms: two-piece distributions (Fernández and Steel, 1998a; Mudholkar and Hutson, 2000; Arellano-Valle et al., 2005) and skew-symmetric distributions (Wang et al., 2004). We propose noninformative benchmark priors and present mild conditions for the existence of the posterior distribution of θ. In Section 5.3, we study the case where X and Y are dependent random variables with skewed marginal distributions. Dependencies between X and Y are modelled using a Gaussian copula. Exploiting the interpretability of the parameters, we provide "vague" proper priors in this context. In Section 5.4, we show that the Bayesian models presented here can be used in the presence of set observations. Finally, Section 5.5 illustrates the use of these models using simulated and real data sets.

5.2 Independent Case

In this section, we present Bayesian models to conduct inference on $\theta = \mathbb{P}(X < Y)$ in the case where X and Y are independent variables with densities $f_1(\cdot; \boldsymbol{\xi_1})$ and $f_2(\cdot; \boldsymbol{\xi_2})$, respectively. Cumulative distribution functions are denoted with the corresponding uppercase letters throughout. We focus on the case where f_1 and f_2 are skewed distributions and we also present conditions for the existence of the posterior distribution of θ under the use of improper benchmark priors.

If we adopt a product prior structure

$$P_{(\boldsymbol{\xi_1},\boldsymbol{\xi_2})} \propto P_{\boldsymbol{\xi_1}} \times P_{\boldsymbol{\xi_2}}, \tag{5.1}$$

where $P_{\boldsymbol{\xi_1}}$ and $P_{\boldsymbol{\xi_2}}$ are priors such that the corresponding posteriors are well-defined, then the posterior distribution of θ is well-defined as shown in the next result.

Remark 1 *Let X and Y be two independent random variables with distributions $f_1(\cdot; \boldsymbol{\xi_1})$ and $f_2(\cdot; \boldsymbol{\xi}_2)$, respectively. Let $\mathbf{x} = (x_1, \dots, x_{n_1})$ and $\mathbf{y} = (y_1, \dots, y_{n_2})$ be two independent samples from X and Y. Then, the posterior distribution of θ, using the product prior structure (5.1), is proper if the corresponding posteriors of $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ are proper.*

Proof. *See Appendix.*

Examples of this are the use of the Jeffreys prior of $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ in a normal or exponential sampling model as in Ventura and Racugno (2011), and the use of reference priors for $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ in a Weibull sampling model as studied in Sun et al. (1998). In the following sections, we study the cases where the marginal distributions of X and Y belong to the family of skewed scale mixtures of normals obtained by two different skewing mechanisms. Let us recall that a density s corresponds to a scale mixture of normals if it can be written as

$$s(x;\nu) = \int_0^\infty \tau^{1/2}\phi(\tau^{1/2}x)dP_{\tau|\nu},$$

where ϕ is the standard normal density and $P_{\tau|\nu}$ is a mixing distribution on $\mathbb{R}_+$. This class is quite wide and covers, for example, Student-t, symmetric stable, exponential power and hyperbolic distributions (see Fernández and Steel, 2000 for a more complete overview).

5.2.1 Two-Piece Marginals

Let s_1 and s_2 be two symmetric densities with support on $\mathbb{R}$, location parameters $\mu_j \in \mathbb{R}$ and scale parameters $\sigma_j \in \mathbb{R}^+$, $j = 1, 2$ respectively. Let X and Y be two independent continuous random variables with densities given respectively by (Arellano-Valle et al., 2005)

$$\begin{aligned} f_1(x;\mu_1,\sigma_1,\gamma_1) &= \frac{2}{\sigma_1[a(\gamma_1)+b(\gamma_1)]} \\ &\times \left[s_1\left(\frac{x-\mu_1}{\sigma_1 b(\gamma_1)}\right) I_{(-\infty,\mu_1)}(x) + s_1\left(\frac{x-\mu_1}{\sigma_1 a(\gamma_1)}\right) I_{[\mu_1,\infty)}(x)\right], \\ f_2(y;\mu_2,\sigma_2,\gamma_2) &= \frac{2}{\sigma_2[a(\gamma_2)+b(\gamma_2)]} \\ &\times \left[s_2\left(\frac{y-\mu_2}{\sigma_2 b(\gamma_2)}\right) I_{(-\infty,\mu_2)}(y) + s_2\left(\frac{y-\mu_2}{\sigma_2 a(\gamma_2)}\right) I_{[\mu_2,\infty)}(y)\right] \end{aligned} \quad (5.2)$$

where $\gamma_j \in \Gamma$ and Γ depends on the choice of $\{a(\cdot), b(\cdot)\}$ where $a(\cdot)$ and $b(\cdot)$ are positive and differentiable functions. The main examples found in the literature are $\{a(\gamma), b(\gamma)\} = \{\gamma, 1/\gamma\}$, $\gamma > 0$ (Fernández and Steel, 1998a) and $\{a(\gamma), b(\gamma)\} = \{1-\gamma, 1+\gamma\}$, $\gamma \in (-1, 1)$ (Mudholkar and Hutson, 2000). The densities f_1 and f_2 can be interpreted as skewed versions of s_1 and s_2, and are often called "two-piece" distributions. If we measure skewness using the measure in Arnold and Groeneveld (1995) (which is defined as one minus twice the probability mass to the left of the mode and takes values in $[-1, 1]$), Rubio and Steel (2011b) find that for these distributions this skewness measure becomes

$$AG = AG(\gamma_j) = \frac{a(\gamma_j) - b(\gamma_j)}{a(\gamma_j) + b(\gamma_j)}, \;\; j = 1, 2.$$

Therefore, we can see that γ_j controls the allocation of mass each side of the mode of the transformed distribution. This result lets us interpret the parameter γ_j as a skewness parameter for the typical choices of $\{a(\cdot), b(\cdot)\}$ found in the literature.

For the purpose of conducting Bayesian inference for the parameter $\theta = \mathbb{P}(X < Y)$ we consider the priors

$$p(\mu_j, \sigma_j, \gamma_j; \alpha_j, \beta_j) \propto \frac{1}{\sigma_j} \frac{|a'(\gamma_j)b(\gamma_j) - a(\gamma_j)b'(\gamma_j)|}{[a(\gamma_j) + b(\gamma_j)]^{\alpha_j+\beta_j}} a(\gamma_j)^{\alpha_j-1} b(\gamma_j)^{\beta_j-1}, \;\; j = 1, 2. \tag{5.3}$$

The structure of these priors is the product of the independence Jeffreys prior for a symmetric location-scale model and a Beta(α_j, β_j) distribution on the parameter $(AG(\gamma_j)+1)/2$ (Rubio and Steel, 2011b). Note that if $\alpha_j = \beta_j = 1$, then the latter prior is equivalent to setting a uniform prior over the measure of skewness AG. This prior structure was proposed in Rubio and Steel (2011b) as a modification of the independence Jeffreys prior for two-piece location-scale models with the aim of producing a proper posterior for a wider range of sampling models than the original one. They also show through a simulation study that the coverage of the credibility intervals obtained with this prior is reasonably close to the nominal value. Conditions for the existence of the posterior distribution of θ using this prior are given in the following corollary.

Corollary 9 *Let* $\mathbf{x} = (x_1, \dots, x_{n_1})$ *and* $\mathbf{y} = (y_1, \dots, y_{n_2})$ *be two independent samples from the models in* (5.2) *and* (5.3)*, where* s_1 *and* s_2 *are scale mixtures of normals. Then,*

(i) The posterior distribution of θ *is proper for any parameterisation* $\{a(\cdot), b(\cdot)\}$ *if* $n_1, n_2 \geq 2$ *and all the observations are different.*

(ii) Suppose that the samples $\mathbf{x}$ *and* $\mathbf{y}$ *contain repeated observations. Let* k_1 *be the largest number of observations with the same value in* $\mathbf{x}$ *and let* k_2 *be the largest number of repeated observations in* $\mathbf{y}$. *If* $1 < k_1 < n_1$ *and* $1 < k_2 < n_2$*, then the posterior of* θ *is proper if and only if the mixing probabilities of* s_1 *and* s_2 *satisfy*

$$\int_{0<\tau_1\leq\cdots\leq\tau_{n_j}<\infty} \tau_{n_j-k_j}^{-(n_j-2)/2} \prod_{i\neq n_j-k_j, n_j} \tau_i^{1/2} dP_{(\tau_1,\dots,\tau_{n_j})} < \infty, \;\; j = 1, 2. \tag{5.4}$$

In the case of a two-piece normal sampling model, it suffices to have two different observations in each sample.

Proof. *(i) is a consequence of Remark 1 above and Theorem 6 from Rubio and Steel (2011b). (ii) follows from the proof of Theorem 6 from Rubio and Steel (2011b) and Theorem 2 from Fernández and Steel (1998b).*

5.2.2 Skew-Symmetric Marginals

We now consider the case where X and Y are independent random variables with skew-symmetric distributions as in Wang et al. (2004). Let s_1 and s_2 be two symmetric densities with support on $\mathbb{R}$, location parameters $\mu_j \in \mathbb{R}$, scale parameters $\sigma_j \in \mathbb{R}^+$, $j = 1, 2$ respectively, and define

$$\begin{aligned} f_1(x; \mu_1, \sigma_1, \pi_1) &= \frac{2}{\sigma_1} s_1\left(\frac{x-\mu_1}{\sigma_1}\right) \pi_1\left(\frac{x-\mu_1}{\sigma_1}\right), \\ f_2(y; \mu_2, \sigma_2, \pi_2) &= \frac{2}{\sigma_2} s_2\left(\frac{y-\mu_2}{\sigma_2}\right) \pi_2\left(\frac{y-\mu_2}{\sigma_2}\right), \end{aligned} \tag{5.5}$$

where $\pi_j(\cdot)$ are functions that satisfy $0 \leq \pi_j(x) \leq 1$ and $\pi_j(-x) = 1 - \pi_j(x)$. We use parametric skewing functions $\pi_j(\cdot; \lambda_j)$, $\lambda_j \in \Lambda_j$, and adopt the prior structure

$$p(\mu_j, \sigma_j, \lambda_j) \propto \sigma_j^{-1} p(\lambda_j), \;\; j = 1, 2, \tag{5.6}$$

where $p(\lambda_j)$ is an integrable function over Λ_j. The structure of these priors is again the product of the independence Jeffreys prior for a symmetric location-scale model and a prior distribution on the skewness parameter λ_j. This prior can be interpreted as an extension of the reference prior of $(\lambda_j, \mu_j, \sigma_j)$ for the skew–normal case calculated in Liseo and Loperfido (2006), which turns out to have this product structure. Bayes and Branco (2007) show that, in the skew-normal case, this prior produces reasonable coverage probabilities under a certain choice of $p_{\lambda_j}(\lambda_j)$ detailed below. Conditions for the existence of the posterior distribution of θ using the prior (5.6) are given in the following corollary.

Corollary 10 *Let* $\mathbf{x} = (x_1, \dots, x_{n_1})$ *and* $\mathbf{y} = (y_1, \dots, y_{n_2})$ *be two independent samples from the model* $(5.5) - (5.6)$*, where* s_1 *and* s_2 *are scale mixtures of normals. Then*

(i) The posterior distribution of θ *is proper if* $n_1, n_2 \geq 2$ *and all the observations are different.*

(ii) Suppose that the samples $\mathbf{x}$ *and* $\mathbf{y}$ *contain repeated observations. Let* k_1 *and* k_2 *be the largest number of observations with the same value in* $\mathbf{x}$ *and* $\mathbf{y}$*, respectively. If* $1 < k_1 < n_1$ *and* $1 < k_2 < n_2$*, then the posterior of* θ *is proper if and only if the mixing probabilities of* s_1 *and* s_2 *satisfy* (5.4)*. In the case of a skew-symmetric normal sampling model, it suffices to have two different observations in each sample.*

Proof. *See Appendix.*

A particular case of model (5.5) is the Azzalini skew-normal (Azzalini, 1985), which is obtained by setting $\pi_j(x;\lambda_j) = \Phi(\lambda_j x)$, $\lambda_j \in \mathbb{R}$, and $s_1 = s_2 = \phi$, where Φ and ϕ are the standard normal CDF and PDF, respectively. This model is frequently used in applications and will be considered for the examples in Section 5.5 together with the prior

$$p(\mu_j, \sigma_j, \lambda_j) \propto {\sigma_j}^{-1} p_J(\lambda_j), \;\; j = 1, 2. \tag{5.7}$$

This prior uses $p_J(\lambda_j)$, which is the Jeffreys prior of λ_j derived in the model without location and scale parameters, and was proposed in Liseo and Loperfido (2006), who also prove existence of the posterior under this prior. Bayes and Branco (2007) show that the Jeffreys prior of λ_j can be approximated by a Student-t distribution with $1/2$ degrees of freedom.

5.3 Dependent Case

In this section, we focus on Bayesian inference for $\theta = \mathbb{P}(X < Y)$ in the case where X and Y are dependent random variables with marginal distributions $f_1(\cdot;\boldsymbol{\xi_1})$ and $f_2(\cdot;\boldsymbol{\xi}_2)$, respectively. We pay special attention to the case where the marginal distributions are skewed and we use a Gaussian copula for modelling dependencies between X and Y. The density of the Gaussian copula is given by

$$\begin{aligned} s(x, y; \boldsymbol{\xi_1}, \boldsymbol{\xi_2}, \rho) &= \frac{1}{\sqrt{1-\rho^2}} \exp\left[-\frac{V^T(R^{-1} - I)V}{2}\right] \\ &\times \;\; f_1(x; \boldsymbol{\xi_1}) f_2(y; \boldsymbol{\xi_2}), \end{aligned} \tag{5.8}$$

where

$$R = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

is a correlation matrix with $\rho \in (-1, 1)$ and $V = (\Phi^{-1}[F_1(x; \boldsymbol{\xi_1})], \Phi^{-1}[F_2(y; \boldsymbol{\xi_2})])^T$. This copula presents some appealing features like being comprehensive, symmetric (in the sense that positive and negative dependence is treated equally) and also that the Spearman's measure of association, $r_\rho \in (-1, 1)$, can be calculated in closed form as (Carta and Steel, 2012)

$$r_\rho = \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right).$$

We adopt a proper prior distribution with independence between ρ, ξ_1 and ξ_2 and density function

$$p(\boldsymbol{\xi_1}, \boldsymbol{\xi_2}, \rho) = p(\boldsymbol{\xi_1})p(\boldsymbol{\xi_2})p(\rho), \tag{5.9}$$

where $p(\boldsymbol{\xi_1})$, $p(\boldsymbol{\xi_2})$ and $p(\rho)$ are probability density functions. Thus, the posterior distribution of θ is well-defined for this Bayesian model. The choice of these priors for the case of two–piece marginals or skew–symmetric marginals is discussed in the next sections.

5.3.1 Two-Piece Marginals

Consider the case where X and Y are dependent random variables with marginal distributions given by (5.2). The dependency between X and Y is modelled with a Gaussian copula as in (5.8). Figure 5.1 shows some contour plots obtained for this copula density using the parameterisation in Mudholkar and Hutson (2000), $\{a(\gamma), b(\gamma)\} = \{1-\gamma, 1+\gamma\}$ and $s_1 = s_2 = \phi$. By appropriately choosing the parameters γ_1, γ_2 and ρ, we can assign a wide range of shapes to the density. The mode of the density is not affected by changes in the parameters, in line with the mode-preserving property of the two-piece skewing mechanism.

For the parameters of this model, we adopt the product prior structure

$$\begin{aligned} p(\mu_1, \mu_2, \sigma_1, \sigma_2, \gamma_1, \gamma_2, \rho) &= p(\mu_1)p(\sigma_1)p(\mu_2)p(\sigma_2)p(\rho) \\ &\times \frac{|a'(\gamma_1)b(\gamma_1) - a(\gamma_1)b'(\gamma_1)|}{[a(\gamma_1)+b(\gamma_1)]^{\alpha_1+\beta_1}} a(\gamma_1)^{\alpha_1-1} b(\gamma_1)^{\beta_1-1} \\ &\times \frac{|a'(\gamma_2)b(\gamma_2) - a(\gamma_2)b'(\gamma_2)|}{[a(\gamma_2)+b(\gamma_2)]^{\alpha_2+\beta_2}} a(\gamma_2)^{\alpha_2-1} b(\gamma_2)^{\beta_2-1}. \end{aligned} \tag{5.10}$$

In order to come up with "vague" or weakly informative proper priors, we consider uniform priors for each of the location parameters (μ_1, μ_2) on a suitable interval. For each of the scale parameters (σ_1, σ_2), we recommend the use of a half-t distribution with scale parameters A_j and ν_j degrees of freedom, $j = 1, 2$. This prior was proposed in Gelman (2006) as a weakly informative prior for this sort of parameters. Of particular interest is the case with $\alpha_j = \beta_j = 1$ in (5.10) together with

$$p(\rho) \propto \frac{1}{1-(\rho/2)^2}, \tag{5.11}$$

which corresponds to $AG \sim U(-1, 1)$ for both marginals and $r_\rho \sim U(-1, 1)$.

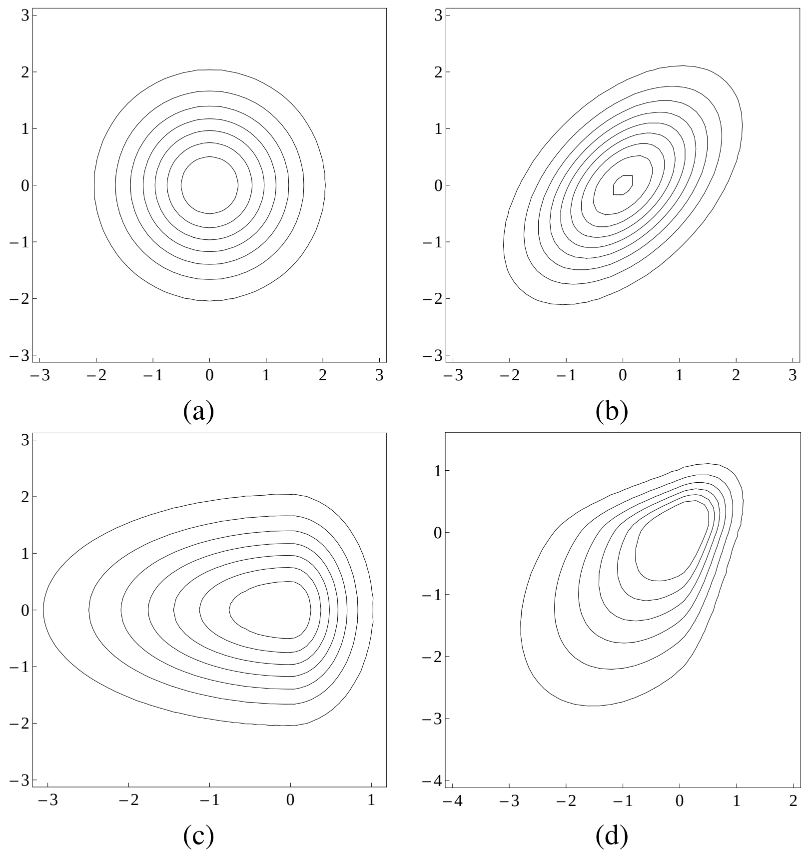


Figure 5.1: Contour plots: two-piece skew-normal marginals with $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ and (a) $\gamma_1 = \gamma_2 = 0$, $\rho = 0$; (b) $\gamma_1 = \gamma_2 = 0$, $\rho = 0.5$; (c) $\gamma_1 = 0.5$, $\gamma_2 = 0$, $\rho = 0$; (d) $\rho = \gamma_1 = \gamma_2 = 0.5$.

5.3.2 Skew-Symmetric Marginals

Here we focus on the case where X and Y are dependent random variables with skew-symmetric marginal distributions (5.5). Figure 5.2 shows some contour plots obtained for the copula density in (5.8) with Azzalini skew-normal marginals. By varying the parameters, it is possible to cover a wide range of shapes, but note that there is a shift of the mode relative to the symmetric case.

For the parameters of this model, we adopt a product structure for the prior

$$p(\mu_1, \mu_2, \sigma_1, \sigma_2, \lambda_1, \lambda_2, \rho) = p(\mu_1)p(\sigma_1)p(\mu_2)p(\sigma_2)p(\lambda_1)p(\lambda_2)p(\rho). \quad (5.12)$$

For $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ we employ the priors described in the previous section. For the skewness parameters (λ_1, λ_2) we employ a Student-t distribution with $1/2$ degrees of freedom as described in Section 5.2.2.

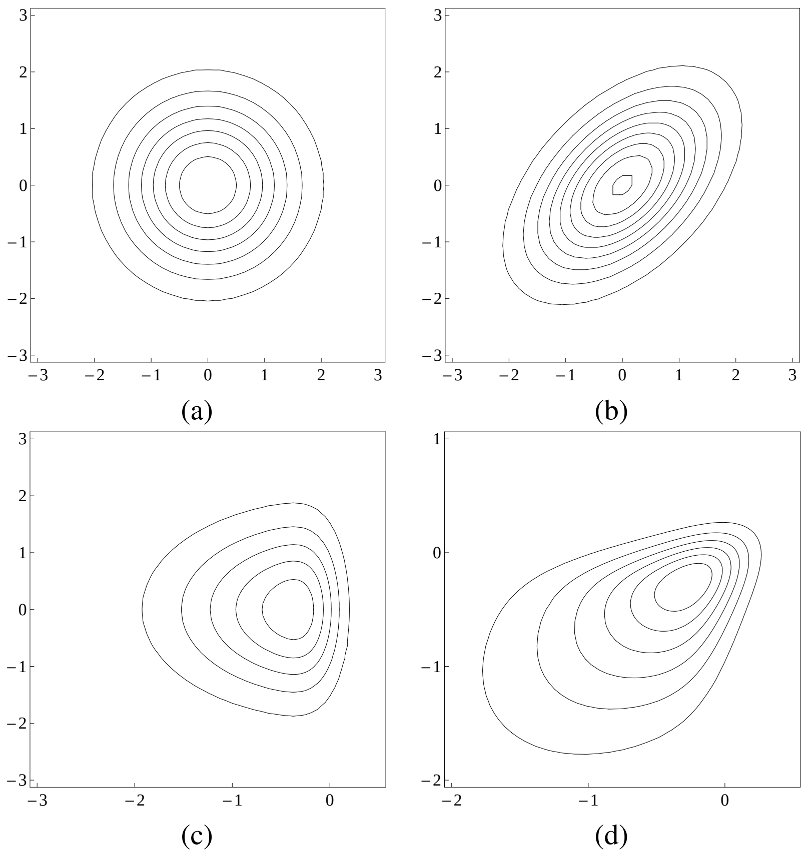


Figure 5.2: Contour plots: Azzalini skew-normal marginals with $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ and (a) $\lambda_1 = \lambda_2 = 0$, $\rho = 0$; (b) $\lambda_1 = \lambda_2 = 0$, $\rho = 0.5$; (c) $\lambda_1 = -5$, $\lambda_2 = 0$, $\rho = 0$; (d) $\lambda_1 = \lambda_2 = -5$, $\rho = 0.5$.

5.4 Set Observations

A common phenomenon in reliability and survival analysis is the presence of set observations under a continuous sampling model. A set observation S is produced when a measurement is recorded as a set of positive probability, *i.e.* $\mathbb{P}[\text{Observing } S] > 0$, where S is a Borel set. In practice, this corresponds to any observation recorded with finite precision, as well as left, right and interval censoring. When the quantitative effect of censoring is significant, this must be formally taken into account in the model (Heitjan, 1989). In addition, the use of set observations allows us to avoid dangerous paradoxes induced by the implicit practice of conditioning on sets of measure zero when using point observations in continuous sampling models (Fernández and Steel, 1998b). In Corollaries 1-2 above, this is reflected in the extra conditions needed in the presence of repeated (point) observations. In the following theorem, conditions for the existence of the posterior distribution of θ using the Bayesian models from Section 5.2 in the context of set (interval) observations are presented.

Theorem 16 *Let* $S_{\mathbf{x}} = (S_1, \dots, S_{n_1})$ *and* $S_{\mathbf{y}} = (S'_1, \dots, S'_{n_2})$ *be two independent samples of set observations from the model* $(5.2) - (5.3)$ *or* $(5.5) - (5.6)$*, where* s_1 *and* s_2 *are scale mixtures of normals. Then, the posterior distribution of* θ *is proper if* $n_1, n_2 \geq 2$ *and there exist two pairs of sets, say* (S_i, S_j) *and* (S'_i, S'_j)*, such that*

$$\begin{aligned} \inf_{x_i \in S_i, x_j \in S_j} |x_i - x_j| &> 0, \\ \inf_{y_i \in S'_i, y_j \in S'_j} |y_i - y_j| &> 0. \end{aligned} \tag{5.13}$$

Proof. *See Appendix.*

Thus, whenever each sample of set observations contains at least two intervals that do not overlap, the posterior distribution of θ is proper. In practice, of course, this is very likely to be satisfied for any samples that we would seriously consider analysing.

For the copula models presented in Section 5.3, the posterior distribution of θ is well-defined in the presence of set observations due to the properness of the priors.

5.5 Examples

In this section, three examples are presented to illustrate the use of the Bayesian models for $\theta = \mathbb{P}(X < Y)$ in different scenarios: independent observations, dependent observations and set observations. Throughout, in order to obtain inferences for θ, we consider the use of the marginal sampling models (5.2) and (5.5) with $s_1 = s_2 = \phi$. In the case of the two-piece marginal we adopt the parameterisation $\{a(\gamma), b(\gamma)\} = \{1 - \gamma, 1 + \gamma\}$, $\gamma \in (-1, 1)$, and use the prior in (5.3) and (5.10) with $\alpha_j = \beta_j = 1$. We compare this model with the Bayesian model with Azzalini skew-normal marginals and the prior in (5.7) and (5.12). For the dependent cases, modelled as in (5.8), we use the prior on ρ in (5.11). Using a Metropolis-Hastings algorithm, a posterior sample of size $10,000$ of the corresponding model parameters was simulated using a burn-in period of $50,000$ iterations and a thinning of 100 iterations. Then, through numerical integration, the corresponding posterior sample of θ was calculated.

5.5.1 Independent Case

Simulated Data

First, we present an example using simulated data which illustrates the importance of taking departures from symmetry into account, particularly in the case where X and Y display quite different skewness properties. Two independent samples of size 50 from the two-piece skew-normal model were drawn with $\mu_i = 10, \sigma_i = 1, i = 1, 2$ and X generated with $\gamma_1 = 0.75$ and Y using $\gamma_2 = -0.75$. Using these data, a posterior sample of θ for the following three Bayesian models was simulated: (i) (5.2)-(5.3), with $\alpha_j = \beta_j = 1$; (ii) (5.5)-(5.7); and (iii) a normal sampling model for X and Y together with the independence Jeffreys prior $p(\mu_1, \mu_2, \sigma_1, \sigma_2) \propto \sigma_1^{-1}\sigma_2^{-1}$. Figure 5.3 shows the posterior distribution of θ for these models. We can observe a clear discrepancy between the inference obtained with the symmetric and the asymmetric sampling models. Properly accounting for skewness centers the inference nicely around the theoretical value (calculated using unidimensional numerical integration) for θ of 0.9646. In addition, both skewed models produce very similar inference about θ.

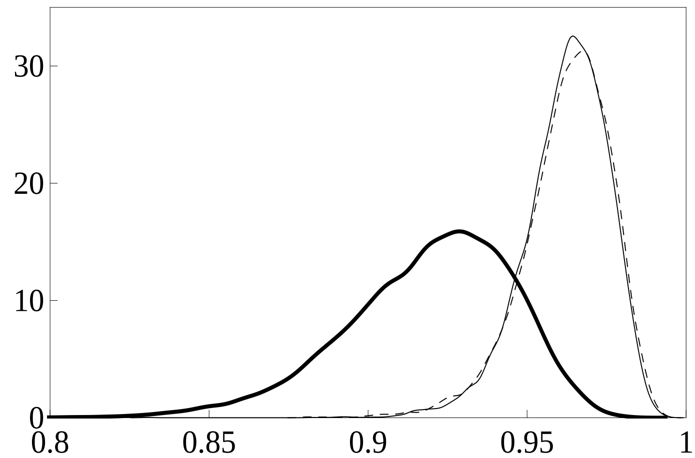


Figure 5.3: Simulated data: posterior distribution of θ, two-piece skew-normal (solid line), Azzalini skew-normal (dashed) and normal (bold).

In the applications with real data, the skewness properties of X and Y are much more similar, and thus the inference on θ is not as crucially affected by allowing for skewness in the marginals. Of course, inference related to the marginals themselves will typically be more sensitive to the modelling of skewness.

Body Measurements

An important goal of forensic studies is to determine the gender of adults given their skeletal remains (Heinz et al., 2003). Therefore, it is important to assess if certain body measurements are informative about the gender. Here we analyse the variable "Chest depth between spine and sternum at nipple level, mid-expiration" from the data set presented in Heinz et al. (2003). This sample consists of 507 measurements taken on physically active adults, 260 females and 247 males. In this case, it seems reasonable to assume independence between the measurements on females and males given that no relationship between the individuals is known. In addition, the histograms in Figure 5.4 suggest departure from symmetry. Figure 5.5 shows the posterior distributions of $\theta = P(\text{female chest depth} < \text{male chest depth})$. This figure indicates that this variable can be informative about the gender given that the posterior of θ assigns most of the mass to values bigger than 0.5. Both models produce similar inferences about θ.

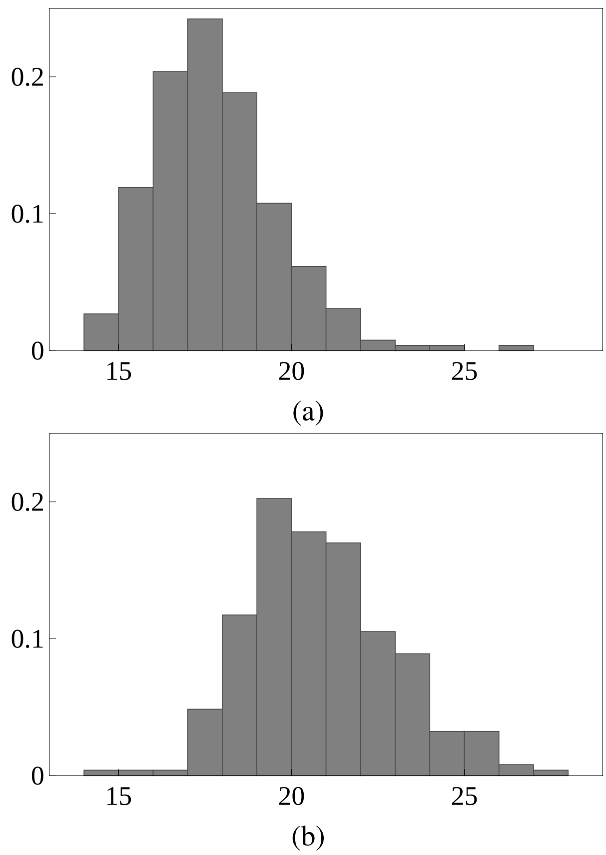


Figure 5.4: Histograms of Chest depth data: (a) females; (b) males.

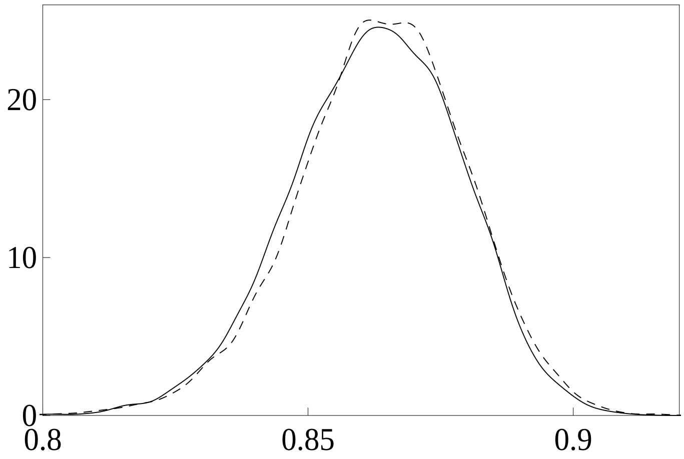


Figure 5.5: Chest depth data: posterior distribution of θ, two-piece skew-normal (solid line) and Azzalini skew-normal (dashed).

5.5.2 Dependent Case

We now analyse the data set presented in Venkatraman and Begg (1996), which contains 72 lesion scores obtained using both, a clinical scheme without a dermoscope (X Test), and a dermoscopic scoring scheme (Y Test). Their main interest is to assess the information provided by the use of the dermoscope. This data set was also considered in Gupta and Peng (2009) using bootstrap and asymptotic confidence intervals but assuming independence between the X Test and the Y Test. This assumption is somewhat restrictive because each pair of observations was measured in the same patient. In fact, the population correlation coefficient is 0.794 and we can observe this positive correlation in the scatter plot in Figure 5.6. Here, we analyse the subset of 51 non-diseased patients (diagnosed using a biopsy) and compare the Bayesian inferences obtained under both assumptions: independence and dependence of the tests. We employ $A_1 = A_2 = \nu_1 = \nu_2 = 1$ for the hyperparameters of the scale parameters in the priors (5.10) and (5.12). For the location parameters we use uniform priors on (-50,50). Figure 5.7 shows the posterior distributions of $\theta = P(\text{Y Test} < \text{X Test})$ for both scenarios. We see that the conclusions are substantially affected by taking the dependence of the variables into account. In contrast, both marginal specifications lead to similar results, as in the previous application. Changing the prior specification by multiplying $A_j, j = 1, 2$ and the boundaries of the uniform priors on μ_1 and μ_2 by a factor 5 or 1/5 does not noticeably affect the results, suggesting a satisfactory amount of prior robustness.

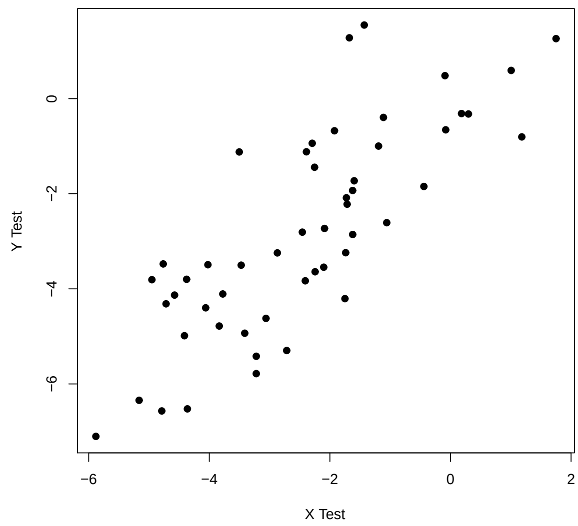


Figure 5.6: Melanoma data: scatter plot.

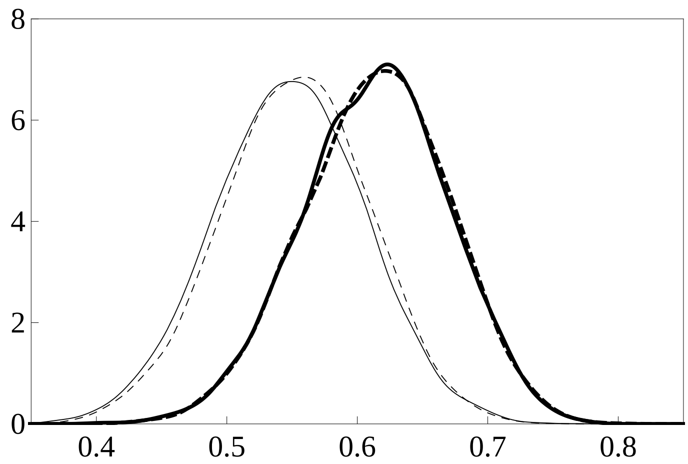


Figure 5.7: Melanoma data: posterior distributions of θ; two-piece skew-normal independent case (solid line), Azzalini skew-normal independent case (dashed), two-piece skew-normal dependent case (bold) and Azzalini skew-normal dependent case (bold dashed).

5.5.3 Set Observations

To illustrate the use of the Bayesian models for θ in the presence of censoring, we consider the breast cancer data set from Finkelstein and Wolfe (1985). This data set contains the times until cosmetic deterioration, determined by evaluation of breast retraction, observed for two treatments (46 observations for the first treatment and 48 observations for the second one): Radiotherapy (R) and Radiotherapy + Chemotherapy (RC). The presence of cosmetic deterioration is observed in between two appointments, so that the observations are recorded as intervals. The assumption of independence between X and Y seems to be reasonable here, but we do take the censoring into account. Since these observations are positive and some of them are close to zero, we analyse the logarithm of the original observations. Figure 5.8 shows the posterior distribution of $\theta = P(\text{R} < \text{RC})$. The posterior mass is clearly concentrated on values smaller than 0.5. This is in line with the conclusion in Finkelstein and Wolfe (1985) that the group receiving both radiotherapy and chemotherapy experiences an earlier cosmetic deterioration.

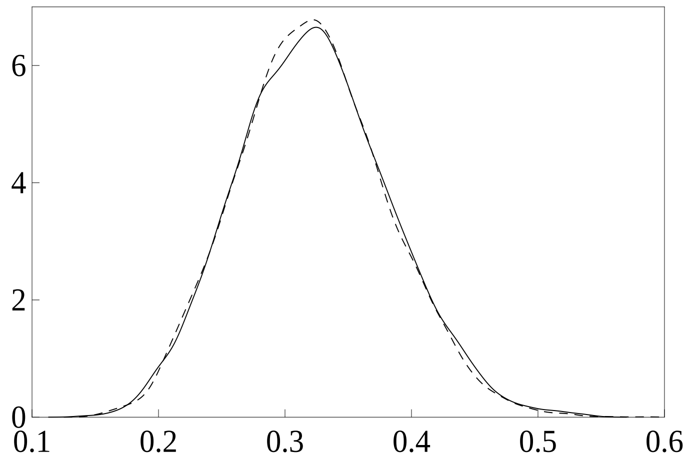


Figure 5.8: Breast cancer data: posterior distributions of θ; two-piece skew-normal model (solid line), Azzalini skew-normal model (dashed).

5.6 Conclusions

We have presented Bayesian models for the parameter $\theta = P(X < Y)$ in the case where the marginal distributions of X and Y belong to the family of skewed scale mixtures of normals. In general, the Bayesian approach overcomes the classical issue regarding the need for an explicit transformation involving this parameter of interest. This allows us to study this problem in more complex scenarios such as the case where X and Y are dependent variables and the context of set observations. Section 5.5 illustrates, through different examples using simulated and real data sets, the relevance of including these assumptions into the model.

Despite the similarities of the inference using two-piece marginals and skew-symmetric marginals observed in the examples, simulating from the posterior distribution of θ using two-piece distributions tends to be easier than with skew-symmetric distributions. Some possible reasons causing this difficulty are the following: (i) in many cases, the likelihood surface of the parameters of skew–symmetric distributions contains a completely flat ridge, and/or the profile likelihood of the parameters can be rather flat (Arnold et al., 1993; Pewsey, 2000); and (ii) the profile likelihood of the shape parameter λ of some skew–symmetric models (such as the Azzalini's skew–normal) has a stationary point at $\lambda = 0$, which has led to some authors to suggest using alternative estimations methods to MLE (Liseo and Loperfido, 2006; Ley and Paindaveine, 2010c). Therefore, when a heavy–tailed prior is employed for the parameters of skew–symmetric models, such as those employed in this chapter, the resulting posterior distribution is often heavy–tailed as well. Due to this, it is usually necessary to use either an adaptive MCMC algorithm or to employ heavy–tailed proposals in order to properly sample from the corresponding tails (see Appendix G.3). We refer the reader to Jarner and Roberts (2007) for a complete study of the use of MCMC methods in contexts with heavy–tailed target distributions.

Finally, we mention two natural directions in which the results presented here can be extended. Firstly, Remark 1 can immediately be applied to contexts with different marginal distributional assumptions. Secondly, we can consider the use of other bivariate copulas (e.g. Archimedean copulas) for modelling dependencies between X and Y.

Chapter 6

Bayesian Modelling of Skewness and Kurtosis With Two-Piece Scale and Shape Transformations

> "In short, our gentleman became so immersed in his reading that he spent whole nights from sundown to sunup and his days from dawn to dusk in poring over his books, until, finally, from so little sleeping and so much reading, his brain dried up and he went completely out of his mind."
> Miguel de Cervantes Saavedra,
> *The Ingenious Gentleman Don Quixote of La Mancha*.

In this chapter we introduce the double two-piece transformation defined on the family of unimodal symmetric continuous distributions containing a shape parameter. The distributions obtained with this transformation contain five interpretable parameters that control the mode, scale and shape in each direction. Symmetric and asymmetric subfamilies of this class of transformations are presented as well as some useful parameterisations. We propose an interpretable scale-and-location invariant benchmark prior for these models and investigate conditions for the existence of the corresponding posterior distribution. Although we focus on the univariate case, we also present two possible multivariate extensions of these transformations.

6.1 Introduction

In the theory of statistical distributions, skewness and kurtosis are features of interest since they provide information about the shape of a distribution. Definitions and quantitative measures of these features have been widely discussed in the statistical literature (see e.g. van Zwet, 1964; Oja, 1981; Groeneveld and Meeden, 1984; Arnold and Groeneveld, 1995; Groeneveld, 1998; Critchley and Jones, 2008).

Distributions containing parameters that control skewness and/or kurtosis have received great attention since they can be used to construct robust models. Although there is no unique way for constructing this sort of flexible distributions, a popular method for doing so consists of adding parameters to a known, typically symmetric, distribution. Transformations that include a parameter that controls skewness are usually referred as "skewing mechanisms" (Ferreira and Steel, 2006; Ley and Paindaveine, 2010a) while those that add a kurtosis parameter have been called "elongations" (Fischer and Klein, 2004), due to the effect produced on the shoulders and the tails of the distributions. Some examples of skewing mechanisms can be found in Azzalini (1985), Fernández and Steel (1998a), and Gupta and Gupta (2008). Examples of elongations can be found in Tukey (1960), Tukey (1977), Haynes et al. (1997), Fischer and Klein (2004), and Klein and Fischer (2006a). A third class of transformations consists of those that contain two parameters that are used for modelling skewness and kurtosis jointly. Some members of this class are the Johnson SU family (Johnson, 1949); Tukey-type transformations such as the g-and-h transformation and the LambertW transformation (Tukey, 1977; Martinez and Iglewicz, 1984; Goerg, 2011), and the sinh-arcsinh transformation (Jones and Pewsey, 2009). These sorts of transformations are typically, but not exclusively, applied to the normal distribution. Alternatively, distributions that can account for skewness and kurtosis can be obtained by introducing skewness into a symmetric distribution that already contains a shape parameter. Examples of distributions obtained by this method are skew-t distributions (Fernández and Steel, 1998a; Azzalini and Capitanio, 2003; Jones and Faddy, 2003; Aas and Haff, 2006; Rosco et al., 2011), and skew-Exponential power distributions (Azzalini, 1986; Fernández et al., 1995). Other distributions containing shape and skewness parameters have been proposed in different contexts such as the hyperbolic distribution (Barndorff-Nielsen, 1977) and the $\alpha-$stable family of distributions. With the exception of the so called "two–piece" transformation (Fernández and Steel, 1998a; Arellano-Valle et al., 2005), the aforementioned transformations produce distributions with different shapes and/or different tail behaviour in each direction.

We introduce a generalisation of the two-piece transformation (Fernández and Steel, 1998a; Arellano-Valle et al., 2005) defined on the family of unimodal, continuous and symmetric distributions that contain a shape parameter. This generalisation consists of using different scale and shape parameters on each side of the mode. We denote this transformation "Double two-piece" (DTP). The resulting distributions contain 5 interpretable parameters that control the mode, scale and shape in each direction. Our proposal transformation contains the original two-piece transformation as a subclass as well as a new class of transformations that can be used for inducing skewness by only varying the shape of the distribution on each side of the mode. This transformation can also be seen as a generalisation of the method proposed in Zhu and Galbraith (2010) for producing a generalised asymmetric Student-t distribution.

The material is organised as follows. In Section 6.2, we introduce the DTP transformation and discuss some of its properties as well as two interesting subfamilies. We present a discussion about the nature of the asymmetry induced by these transformations as well as some useful reparameterisations. We also discuss two possible multivariate extensions of the proposed transformations. In Section 6.3 we propose a scale-and-location invariant "benchmark prior" for the proposed models and conditions for the existence of the corresponding posterior distribution. In Section 6.4 we explore the use of this Bayesian model together with a sampling model obtained by transforming the Student-t distribution. In Section 6.5 we present four examples using real data. These examples show different scenarios where the data exhibit either similar or different tail behaviour in each direction.

6.2 Two-Piece Scale and Shape Transformations

Let $\mathcal{F}$ be the family of continuous, unimodal, symmetric densities $f(\cdot;\mu,\sigma,\delta)$ with mode and location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma \in \mathbb{R}_+$, and shape parameter $\delta \in \Delta \subset \mathbb{R}$. Recall that a shape parameter is defined as (Everitt, 2002)

Definition 6 Shape parameter *is a general term for a parameter of a probability distribution function that determines the* shape *(in a sense distinct from location and scale) of the distribution within a family of shapes associated with a specified type of variable.*

This definition basically indicates that a shape parameter is neither a location parameter nor a scale parameter.

Denote $f(x;\mu,\sigma,\delta) = \dfrac{1}{\sigma} f\left(\dfrac{x-\mu}{\sigma};\delta\right)$. Distribution functions will be denoted with the corresponding uppercase letters. Consider the two-piece density constructed of $f(x;\mu,\sigma_1,\delta_1)$ truncated at $(-\infty,\mu)$ and $f(x;\mu,\sigma_2,\delta_2)$ truncated at $[\mu,\infty)$:

$$\begin{aligned} s(x;\mu,\sigma_1,\sigma_2,\delta_1,\delta_2) &= \frac{2\varepsilon}{\sigma_1} f\left(\frac{x-\mu}{\sigma_1};\delta_1\right) I(x<\mu) \\ &+ \frac{2(1-\varepsilon)}{\sigma_2} f\left(\frac{x-\mu}{\sigma_2};\delta_2\right) I(x\geq\mu), \end{aligned} \tag{6.1}$$

where

$$\varepsilon = \frac{\sigma_1 f(0;\delta_2)}{\sigma_1 f(0;\delta_2) + \sigma_2 f(0;\delta_1)}. \tag{6.2}$$

The corresponding cumulative distribution function (CDF) is then given by

$$\begin{aligned} S(x;\mu,\sigma_1,\sigma_2,\delta_1,\delta_2) &= 2\varepsilon F\left(\frac{x-\mu}{\sigma_1};\delta_1\right) I(x<\mu) \\ &+ \left\{\varepsilon + (1-\varepsilon)\left[2F\left(\frac{x-\mu}{\sigma_2};\delta_2\right) - 1\right]\right\} I(x\geq\mu). \end{aligned} \tag{6.3}$$

By construction, the density (6.1) is continuous, unimodal with mode at μ, and the allocation of mass on each side of its mode is given by (6.2). This transformation preserves the ease of use of the original distribution f and allows s to have different shapes in each direction, dictated by the parameters δ_1 and δ_2. The family $\mathcal{F}$, on which the proposed transformation is defined, contains some important distributions such as the symmetric Johnson-SU distribution (Johnson, 1949), the symmetric sinh-arcsinh distribution (Jones and Pewsey, 2009), and the family of scale mixtures of normals which includes, for instance, the Student-t distribution, the exponential power distribution, the symmetric hyperbolic distribution (Barndorff-Nielsen, 1977), and the symmetric $\alpha-$stable family. Expressions for the density of some of these models are presented below. The shape parameter, $\delta > 0$, in all these models has been interpreted as a kurtosis parameter.

- The symmetric Johnson-SU distribution (Johnson, 1949).

$$f(x;\mu,\sigma,\delta) = \frac{\delta}{\sigma}\phi\left[\delta \operatorname{arcsinh}\left(\frac{x-\mu}{\sigma}\right)\right]\left(1+\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{1}{2}}.$$

- The symmetric sinh-arcsinh distribution (Jones and Pewsey, 2009).

$$f(x;\mu,\sigma,\delta) = \frac{\delta}{\sigma}\phi\left[\sinh\left(\delta \operatorname{arcsinh}\left(\frac{x-\mu}{\sigma}\right)\right)\right]\frac{\cosh\left(\delta \operatorname{arcsinh}\left(\frac{x-\mu}{\sigma}\right)\right)}{\sqrt{1+\left(\frac{x-\mu}{\sigma}\right)^2}}.$$

- The symmetric hyperbolic distribution (Barndorff-Nielsen, 1977).

$$f(x;\mu,\sigma,\delta) = \frac{1}{2\sigma\delta K_1(\delta)} \exp\left[-\sqrt{\delta^2 + \left(\frac{x-\mu}{\sigma}\right)}\right],$$

where K_1 is the modified Bessel of the second kind with index 1.

- The Student-t distribution.

$$f(x;\mu,\sigma,\delta) = \frac{\Gamma\left(\frac{\delta+1}{2}\right)}{\sigma\sqrt{\pi\delta}\Gamma\left(\frac{\delta}{2}\right)} \left(1 + \frac{\left(\frac{x-\mu}{\sigma}\right)^2}{\delta}\right)^{-\frac{\delta+1}{2}},$$

where $\Gamma(\cdot)$ is the gamma function.

- The exponential power distribution.

$$f(x;\mu,\sigma,\delta) = \frac{\delta}{2\sigma\Gamma(1/\delta)} \exp\left(-\frac{|x-\mu|^\delta}{\sigma^\delta}\right),$$

where $\Gamma(\cdot)$ is the gamma function.

Transformation (6.1) preserves the existence of moments, when they exist for both δ_1 and δ_2, since

$$\begin{aligned}\int_{\mathbb{R}} x^r s(x;\mu,\sigma_1,\sigma_2,\delta_1,\delta_2)dx &= 2\varepsilon \int_{-\infty}^{\mu} x^r f(x;\mu,\sigma_1,\delta_1)dx \\ &+ 2(1-\varepsilon)\int_{\mu}^{\infty} x^r f(x;\mu,\sigma_2,\delta_2)dx.\end{aligned}$$

For example, if f in (6.1) is the Student-t density with δ degrees of freedom, then the rth moment of s exists when both $\delta_1, \delta_2 > r$. Figure 6.1 shows the shapes that we can obtain by applying this method to the symmetric sinh-arcsinh distribution.

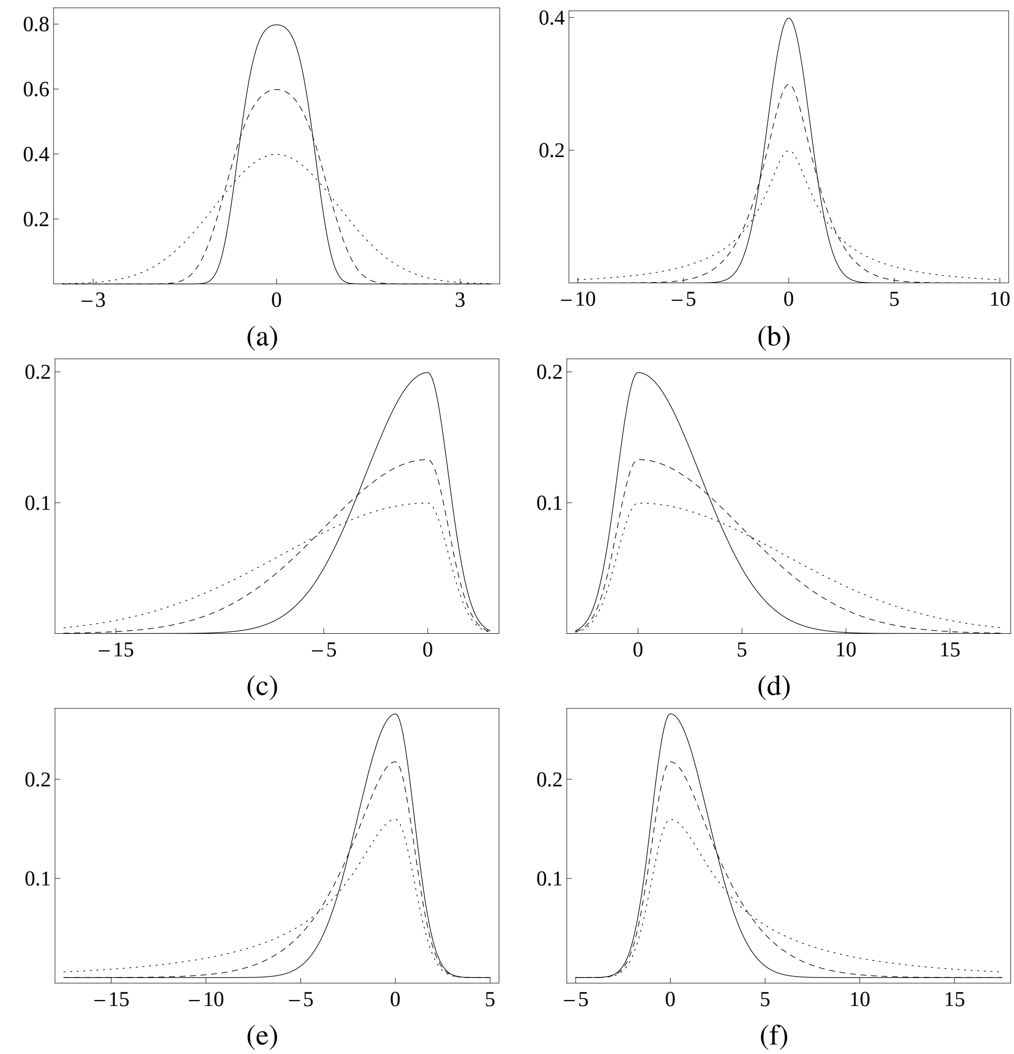


Figure 6.1: DTP sinh-arcsinh distribution with $\mu = 0$ and: (a) $(\sigma_1, \sigma_2) = (1, 1)$, $\delta_1 = \delta_2 = 2, 1.5, 1$; (b) $(\sigma_1, \sigma_2) = (1, 1)$, $\delta_1 = \delta_2 = 1, 0.75, 0.5$; (c) $\sigma_1 = 3, 5, 7$, $\sigma_2 = \delta_1 = \delta_2 = 1$; (d) $\sigma_1 = 1$, $\sigma_2 = 3, 5, 7$, $\delta_1 = \delta_2 = 2$; (e) $\sigma_1 = 2$, $\sigma_2 = 1$, $\delta_1 = 1, 0.75, 0.5$, $\delta_2 = 1$; (f) $\sigma_1 = 1$, $\sigma_2 = 2$, $\delta_1 = 1$, $\delta_2 = 1, 0.75, 0.5$.

6.2.1 4-parameter Asymmetric Subfamilies

Two-Piece Scale Transformation

The DTP family of transformations naturally includes the original two–piece transformation by setting the condition $\delta_1 = \delta_2 = \delta$ in (6.1)

$$s(x;\mu,\sigma_1,\sigma_2,\delta) = \frac{2}{\sigma_1+\sigma_2}\left[f\left(\frac{x-\mu}{\sigma_1};\delta\right)I(x<\mu) + f\left(\frac{x-\mu}{\sigma_2};\delta\right)I(x\geq\mu)\right]. \tag{6.4}$$

This subfamily will be denoted "two–piece scale" (TPSC). The cases where $f(\cdot;\delta)$ is a Student t distribution or a exponential power distribution have already been analysed (Fernández et al., 1995) but, as far as we are aware, the cases where f is a symmetric hyperbolic distribution, or a symmetric Johnson-SU distribution, or a symmetric sinh-arcsinh distribution have not been studied yet. Expressions for the densities of these models can be obtained by using the corresponding symmetric density $f(\cdot;\delta)$ in (6.4).

Two-Piece Shape Transformation

An alternative subfamily can be obtained by fixing $\sigma_1 = \sigma_2 = \sigma$ in (6.1)

$$\begin{aligned} s(x;\mu,\sigma,\delta_1,\delta_2) &= \frac{2\varepsilon}{\sigma}f\left(\frac{x-\mu}{\sigma};\delta_1\right)I(x<\mu) \\ &+ \frac{2(1-\varepsilon)}{\sigma}f\left(\frac{x-\mu}{\sigma};\delta_2\right)I(x\geq\mu), \end{aligned} \tag{6.5}$$

where $\varepsilon = \dfrac{f(0;\delta_2)}{f(0;\delta_1)+f(0;\delta_2)}$. This subfamily will be denoted "two–piece shape" (TPSH). This transformation, which has not been studied yet to the best of our knowledge, produces distributions with different shape parameters in each direction and can be used to induce asymmetry whenever the shape parameter δ is identifiable. It is important to notice that, by using the TPSH transformation, skewness can only be introduced if the shape parameters of the original distribution differ in each direction. Related distributions have been studied, for instance, in Jones and Faddy (2003) and Aas and Haff (2006). Figure 6.2 shows two examples of distributions obtained with this transformation.

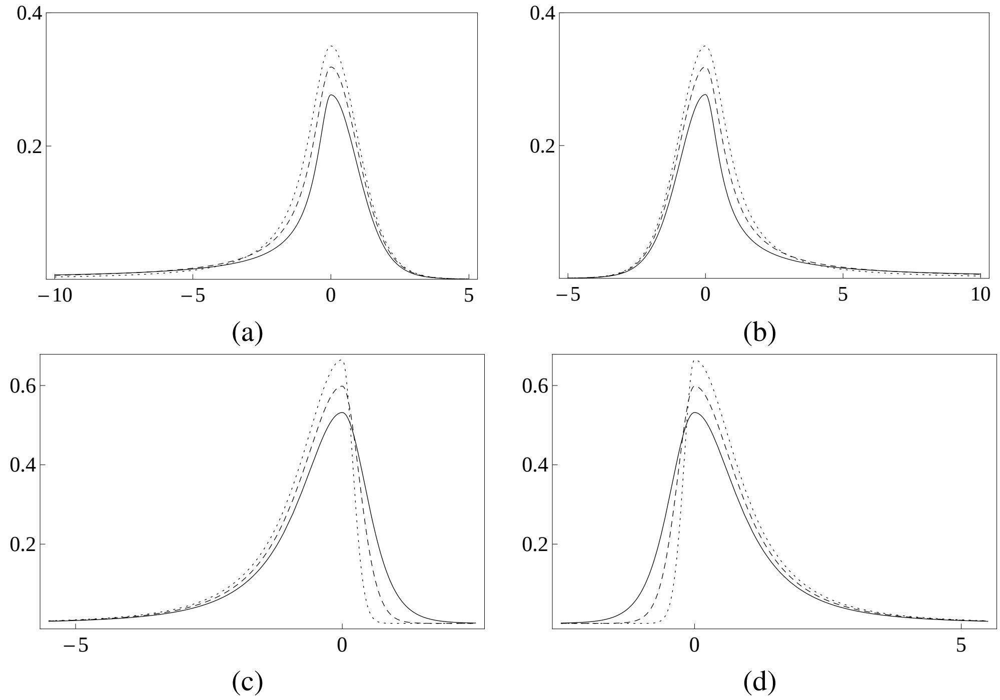


Figure 6.2: TPSH densities with $(\mu, \sigma) = (0, 1)$: (a) TPSH Student-t, $\delta_1 = 0.25, 0.5, 1$, $\delta_2 = 10$; (b) TPSH Student-t, $\delta_1 = 10$, $\delta_2 = 0.25, 0.5, 1$; (c) TPSH Johnson-SU, $\delta_1 = 1$, $\delta_2 = 2, 3, 5$; (d) TPSH Johnson-SU, $\delta_1 = 2, 3, 5$, $\delta_2 = 1$.

6.2.2 Understanding the Skewing Mechanism Induced by the Proposed Transformations

In order to provide more insight into the transformation (6.1), we analyse the TPSC and TPSH families of transformations separately. In the TPSC family, asymmetry is produced by varying the scale parameters on each side of the mode. This simply reallocates the mass of the density while preserving the tail behaviour in each direction. Since the nature of the asymmetry induced by the TPSC transformation is clear, we now focus on the study of TPSH transformations. For this purpose we employ two measures of asymmetry, the Critchley-Jones (CJ) functional asymmetry measure (Critchley and Jones, 2008) and the Arnold-Groeneveld (AG) scalar measure of skewness (Arnold and Groeneveld, 1995). The CJ functional of asymmetry measures discrepancies of points located on each side of the mode $(x_L(p), x_R(p))$ of a density f such that $f(x_L(p)) = f(x_R(p)) = pf(\text{mode})$, $p \in (0,1)$. This is defined as follows

$$\text{CJ}(p) = \frac{x_R(p) - 2\text{mode} + x_L(p)}{x_R(p) - x_L(p)}. \tag{6.6}$$

Note that this measure takes values in $(-1, 1)$; negative values of $\text{CJ}(p)$ indicate that the values $x_L(p)$ are further from the mode than the values $x_R(p)$. An analogous interpretation applies to positive values. The AG measure of skewness is defined as $1 - 2F(\text{mode})$, where F is the CDF associated to f. This measure also takes values in $(-1, 1)$; negative values of AG are associated to left-skewness and positive values correspond to right-skewness. For models of type (6.1) these quantities are easy to calculate since $\text{AG}(\sigma_1, \sigma_2, \delta_1, \delta_2) = 1 - 2\varepsilon$, and

$$\text{CJ}(p) = \frac{\sigma_2 f_R^{-1}(pf(0;\delta_2);\delta_2) + \sigma_1 f_L^{-1}(pf(0;\delta_1);\delta_1)}{\sigma_2 f_R^{-1}(pf(0;\delta_2);\delta_2) - \sigma_1 f_L^{-1}(pf(0;\delta_1);\delta_1)}, \tag{6.7}$$

where $f_L^{-1}(\cdot;\delta)$ and $f_R^{-1}(\cdot;\delta)$ represent the negative and positive inverse of $f(\cdot;\delta)$, respectively. Note also that $\text{CJ}(p) = \text{AG}$ when $\delta_1 = \delta_2$ for every $p \in (0,1)$.

Figure 6.3 shows some examples of (6.7) with distributions obtained using the TPSH transformation with parameters and AG indicated in Table 6.1. Figure 6.3a shows an example where $\text{CJ}(p)$ changes sign in cases where AG is either positive or negative. This means that the mass cumulated on each side of the mode of the density is different and that the behaviour of the distance of the points $(x_L(p), x_R(p))$ to the mode varies from the tails to the mode of the density as a consequence of the different shapes. Figures 6.3b and 6.3c, on the other hand, show densities with different mass on each side of the mode but with constant sign $\mathrm{CJ}(p)$. Figure 6.3d provides an interesting example where the original transformation contains a shape parameter that hardly modifies the shape of the density and therefore $\mathrm{CJ}(p)$ is virtually constant for all values of $p \in (0,1)$.

These examples illustrate differences between the families TPSC and TPSH. The TPSC transformation basically modifies the mass cumulated on each side of the mode while preserving the shape. The TPSH transformation can also be used to produce asymmetric distributions by modifying the shape in each direction. However, when the TPSH transformation is applied to some distributions $f \in \mathcal{F}$, such as the Student-t, a substantially different shape in each direction is required to accommodate even moderate amounts of skewness in terms of AG (see Table 6.1). For these reasons, we believe these two families of transformations are complementary and that their combination through (6.1) is an appealing idea for obtaining distributions that can capture different features.

TPSH Student-t			TPSH sinh-arcsinh			TPSH Johnson-SU			TPSH hyperbolic		
δ_1	δ_2	AG	δ_1	δ_2	AG	δ_1	δ_2	AG	δ_1	δ_2	AG
1/10	10	-0.45	5	1	2/3	5	1	2/3	1	50	0.69
1/2	10	-0.18	5	2	0.43	5	2	0.43	1	10	0.43
1	10	-0.1	1	1/4	3/5	1	1/4	3/5	1	5	0.29
5	10	-0.01	1	1/2	1/3	1	1/2	1/3	2	1	0.11
10	5	0.01	1/2	1	-1/3	1/2	1	-1/3	1	2	-0.11
10	1	0.1	1/4	1	-3/5	1/4	1	-3/5	5	1	-0.29
10	1/2	0.18	2	5	-0.43	2	5	-0.43	10	1	-0.43
10	1/10	0.45	1	5	-2/3	1	5	-2/3	50	1	-0.69

Table 6.1: Parameters used to obtain the functionals in Figure 6.3.

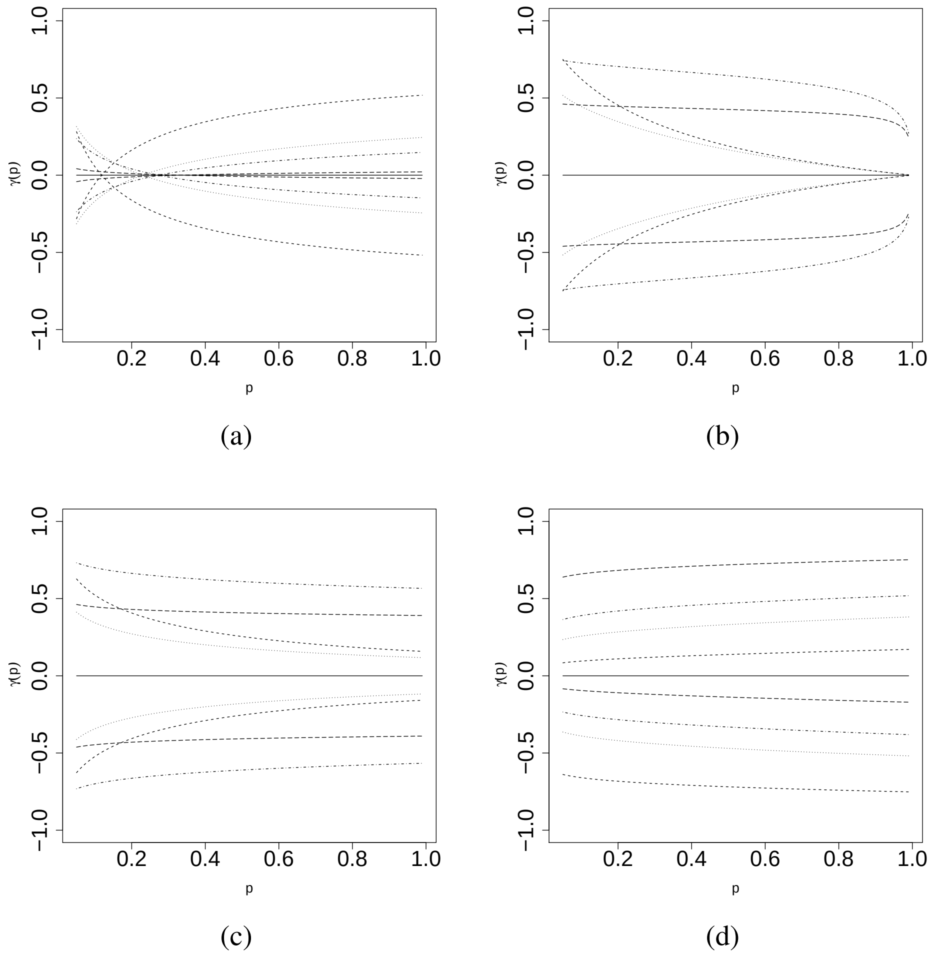


Figure 6.3: Asymmetry functional CJ for: (a) TPSH Student t distribution; (b) TPSH sinh-arcsinh distribution; (c) TPSH Johnson-SU distribution; (d) TPSH hyperbolic distribution.

6.2.3 Some Reparameterisations

For the TPSC family (6.4), Arellano-Valle et al. (2005) propose the reparameterisation $(\mu, \sigma_1, \sigma_2, \delta) \leftrightarrow (\mu, \sigma, \gamma, \delta)$ using the transformation $\sigma_1 = \sigma b(\gamma)$, $\sigma_2 = \sigma a(\gamma)$; where $\{a(\cdot), b(\cdot)\}$ are differentiable functions, $\gamma \in \Gamma$, and the parameter space Γ depends on the choice of $\{a(\cdot), b(\cdot)\}$. The most common choices for $a(\cdot)$ and $b(\cdot)$ correspond to the *inverse scale factors* parameterisation $\{a(\gamma), b(\gamma)\} = \{\gamma, 1/\gamma\}$, $\gamma \in \mathbb{R}_+$ (Fernández and Steel, 1998a), and the $\epsilon-$*skew* parameterisation $\{a(\gamma), b(\gamma)\} = \{1-\gamma, 1+\gamma\}$, $\gamma \in (-1, 1)$ (Mudholkar and Hutson, 2000). Jones and Anaya-Izquierdo (2010) and Rubio and Steel (2011b) analyse the choices of $\{a(\cdot), b(\cdot)\}$ that induce orthogonality between σ and γ. This reparameterisation is also appealing because it allows the interpretation of γ as a skewness parameter since the AG measure of skewness depends only on this parameter.

This reparameterisation can be used in the family of DTP transformations (6.1) as well for inducing orthogonality between σ and γ. Under this reparameterisation, density (6.1) becomes

$$
\begin{aligned}
s(x; \mu, \sigma, \gamma, \delta) &= \frac{2}{\sigma c(\gamma, \delta_1, \delta_2)} \left[f(0; \delta_2) f\left(\frac{x-\mu}{\sigma b(\gamma)}; \delta_1\right) I(x < \mu) \right. \\
&+ \left. f(0; \delta_1) f\left(\frac{x-\mu}{\sigma a(\gamma)}; \delta_2\right) I(x \geq \mu) \right], \qquad (6.8)
\end{aligned}
$$

where $c(\gamma, \delta_1, \delta_2) = b(\gamma) f(0; \delta_2) + a(\gamma) f(0; \delta_1)$. The interpretation of γ in this family is slightly different since the cumulation of mass depends also on the shape parameters (δ_1, δ_2). However, the parameter γ does not modify the shape of s.

Using this reparameterisation we can obtain the "generalized asymmetric Student-t distribution" proposed in Zhu and Galbraith (2010) by taking $f(\cdot; \delta)$ as a Student-t distribution and $\{a(\gamma), b(\gamma)\} = \{\gamma, 1-\gamma\}$, $\gamma \in (0, 1)$. Under the same parameterisation, we can obtain the "generalized asymmetric exponential power distribution" proposed in Zhu and Galbraith (2011) by taking $f(\cdot; \delta)$ as a exponential power distribution.

Unfortunately, for the TPSH family (6.5) there seems to be no obvious reparameterisation that induces parameter orthogonality.

6.2.4 Extensions to the Multivariate Case

DTP, TPSC and TPSH families can be extended to the multivariate case in several ways. For instance, Ferreira and Steel (2007) propose the use of affine transformations to produce a multivariate extension of TPSC models while Rubio and Steel (2013) propose the use of copulas. In a similar fashion, the DTP (and consequently the TPSH) family can be extended to the multivariate scenario as described below. For illustrative purposes we consider the parameterisation in (6.8).

Copulas

The use of copulas with DTP marginals (6.8) is an appealing option since the resulting models are closed under marginalisation. Multivariate models presenting this property were termed "coherent with respect to marginalization" in Sahu et al. (2003). A natural first choice is the Gaussian copula which presents nice properties such as being comprehensive, symmetric (in the sense that positive and negative dependence is treated equally) and allowing for an extension to any dimension. The density of the Gaussian copula with DTP marginals in d dimensions is given by

$$\begin{aligned} s_d(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{R},\boldsymbol{\gamma},\boldsymbol{\delta}) &= \frac{1}{\sqrt{\det \boldsymbol{R}}}\exp\left[-\frac{V^T(\boldsymbol{R}^{-1}-\boldsymbol{I})V}{2}\right] \\ &\times \prod_{j=1}^{d} s(x_j;\mu_j,\sigma_j,\gamma_j,\delta_j), \end{aligned}$$

where $\boldsymbol{x} = (x_1,\dots,x_d) \in \mathbb{R}^d$, $\boldsymbol{\mu} = (\mu_1,\dots,\mu_d) \in \mathbb{R}^d$, $\boldsymbol{\sigma} = (\sigma_1,\dots,\sigma_d) \in \mathbb{R}_+^d$, $\boldsymbol{\gamma} = (\gamma_1,\dots,\gamma_d) \in \Gamma^d$, $\boldsymbol{\delta} = (\delta_1,\dots,\delta_d) \in \Delta^d$, $\boldsymbol{R}$ is a correlation matrix, and $V = (\Phi^{-1}[S(x_1;\mu_1,\sigma_1,\gamma_1,\delta_1)],\dots,\Phi^{-1}[S(x_d;\mu_d,\sigma_d,\gamma_d,\delta_d)])^\top$.

Affine transformations

Let $X_1,\dots,X_d$ be independent random variables with densities as in (6.8) with parameters $(0,1,\gamma_j,\delta_j)$, $j = 1,\dots,d$, and $f \in \mathcal{F}$ be allowed to vary for $j = 1,\dots,d$. Let $X = (X_1,...,X_d)^\top$ and define $Y = \Sigma^\top X + \boldsymbol{\mu}$, where $\boldsymbol{\mu} = (\mu_1,\dots,\mu_d)^\top \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}_+^{d\times d}$ is a non-singular matrix. The PDF of Y is then given by

$$s_d(\boldsymbol{y};\boldsymbol{\mu},\Sigma,\boldsymbol{\gamma},\boldsymbol{\delta}) = |\det\Sigma|^{-1}\prod_{j=1}^{d} s\left[(y-\mu)^\top\Sigma_{\cdot j}^{-1};\gamma_j,\delta_j\right],$$

where $\boldsymbol{y} = (y_1,...,y_d)^\top \in \mathbb{R}^d$, $\boldsymbol{\gamma} = (\gamma_1,\dots,\gamma_d)^\top \in \Gamma^d$, $\boldsymbol{\delta} = (\delta_1,\dots,\delta_d)^\top \in \Delta^d$, and $\Sigma_{\cdot j}^{-1}$ denotes the jth column of Σ^{-1}. The distribution of Y is also unimodal with mode at $\boldsymbol{\mu}$. Expressions for the marginal distributions of this sort of extension have been studied in Pinheiro (2012). They pointed out that this family of multivariate extensions is not closed under marginalisation in general.

6.3 Bayesian Inference

In this section we propose a class of "benchmark" priors for the models studied in Section 6.2 with the parameterisation in (6.8). The proposed prior structure is inspired by the independence Jeffreys prior of the symmetric model, producing a scale-and-location invariant prior. Conditions for the existence of the posterior distribution under the use of this prior are provided for the case where f is a scale mixture of normals. The case where the sample contains repeated observations is covered as well.

Let us recall that a density f with shape parameter δ corresponds to a scale mixture of normals if it can be written as

$$f(x;\delta) = \int_0^\infty \tau^{1/2}\phi(\tau^{1/2}x)dP_{\tau|\delta},$$

where ϕ is the standard normal density and $P_{\tau|\delta}$ is a mixing distribution on $\mathbb{R}_+$. This is a broad class of distributions that include, for instance, the Student-t distribution, the symmetric $\alpha-$stable, the exponential power distribution, and the symmetric hyperbolic distribution (see Fernández and Steel, 2000 for a more complete overview).

As discussed in previous sections, the parameters of a distribution obtained through the TPSC transformation, $(\mu,\sigma,\gamma,\delta)$, can be interpreted as location, scale, skewness and shape, respectively. For this reason we adopt the product prior structure $p(\mu,\sigma,\gamma,\delta) \propto \dfrac{1}{\sigma}p(\gamma)p(\delta)$. The following result provides conditions for the existence of the posterior distribution under the use of this prior structure.

Remark 2 (TPSC family) *Let* $\mathbf{x} = (x_1,...,x_n)$ *be an* i.i.d. *sample from* (6.8) *with* $\delta_1 = \delta_2 = \delta$. *Let* f *be a scale mixture of normals and consider the prior*

$$p(\mu,\sigma,\gamma,\delta) \propto \frac{1}{\sigma}p(\gamma)p(\delta), \tag{6.9}$$

where $p(\gamma)$ *and* $p(\delta)$ *are proper priors. Then*

(i) The posterior distribution of $(\mu,\sigma,\gamma,\delta)$ *is proper if* $n \geq 2$ *and all the observations are different.*

(ii) If $\mathbf{x}$ *contains repeated observations, let* k *be the largest number of observations with the same value in* $\mathbf{x}$ *and* $1 < k < n$, *then the posterior of* $(\mu,\sigma,\gamma,\delta)$ *is proper if and*

only if the mixing probability of f satisfies

$$\int_{0<\tau_1\leq\cdots\leq\tau_n<\infty} \tau_{n-k}^{-(n-2)/2} \prod_{i\neq n-k,n} \tau_i^{1/2} dP_{(\tau_1,\dots,\tau_n)} < \infty. \tag{6.10}$$

In the case of a two-piece Student-t sampling model, (6.10) *is equivalent to*

$$\int_{(k-1)/(n-k)}^{(k-1)/(n-k)+\xi} \frac{p(\delta)}{(n-k)\delta-(k-1)} d\delta < \infty \text{ and } \int_0^{(k-1)/(n-k)} p(\delta)d\delta = 0, \tag{6.11}$$

for all $\xi > 0$.

Proof. *These results were proved in Theorems* 1, 2 *and* 3 *from Fernández and Steel (1998b).*

For the parameters of TPSH models we have that the shape parameters (δ_1, δ_2) control the mass cumulated on each side of the mode as well as the shape. In addition, these parameters are not orthogonal in general. For these reasons we adopt a product prior structure with a joint distribution on (δ_1, δ_2) in order to include priors that can account for dependencies between these parameters. The next result provides conditions for the existence of the posterior distribution using this prior structure.

Theorem 17 (TPSH family) *Let $\mathbf{x} = (x_1, ..., x_n)$ be an* i.i.d. *sample from* (6.8) *with $\sigma_1 = \sigma_2 = \sigma$. Let f be a scale mixture of normals and consider the prior structure*

$$p(\mu, \sigma, \delta_1, \delta_2) \propto \frac{1}{\sigma} p(\delta_1, \delta_2), \tag{6.12}$$

where $p(\delta_1, \delta_2)$ is a proper prior. Then

(i) The posterior distribution of $(\mu, \sigma, \delta_1, \delta_2)$ is proper if $n \geq 2$ and all the observations are different.

(ii) If $\mathbf{x}$ contains repeated observations, let k be the largest number of observations with the same value in $\mathbf{x}$ and $1 < k < n$, then the posterior of $(\mu, \sigma, \delta_1, \delta_2)$ is proper if and only if the mixing probability of f satisfies (6.10).

Proof. *See Appendix*

Finally, in DTP models the parameters $(\gamma, \delta_1, \delta_2)$ jointly control the shape and the mass cumulated in each direction. Then, we adopt a product prior structure with a joint distribution on $(\gamma, \delta_1, \delta_2)$ allowing for the inclusion of prior beliefs on the dependence between these parameters. The following result provides conditions for the existence of the corresponding posterior distribution.

Theorem 18 (DTP family) *Let* $\mathbf{x} = (x_1, ..., x_n)$ *be an independent sample from* (6.8). *Let* f *be a scale mixture of normals and consider the prior structure*

$$p(\mu, \sigma, \gamma, \delta_1, \delta_2) \propto \frac{1}{\sigma} p(\gamma, \delta_1, \delta_2), \tag{6.13}$$

where $p(\gamma, \delta_1, \delta_2)$ *is a proper prior. It follows that*

(i) The posterior distribution of $(\mu, \sigma, \gamma, \delta_1, \delta_2)$ *is proper if* $n \geq 2$ *and all the observations are different.*

(ii) If $\mathbf{x}$ *contains repeated observations, let* k *be the largest number of observations with the same value in* $\mathbf{x}$ *and* $1 < k < n$*, then the posterior of* $(\mu, \sigma, \gamma, \delta_1, \delta_2)$ *is proper if and only if the mixing probability of* f *satisfies* (6.10).

Proof. *See Appendix*

In applications we believe it is reasonable to use a product structure $P_{\gamma,\delta_1,\delta_2} = P_\gamma \times P_{\delta_1,\delta_2}$, given that the parameter γ does not modify the shape of the distribution. This case is of course covered by Theorem 18.

Some possible choices for the priors $p(\gamma)$, $p(\delta)$, and $p(\delta_1, \delta_2)$ involved in the previous results are discussed in the next section for the case where f is a Student-t distribution, used for the examples studied in Section 6.5.

6.4 Priors for Student-t Base Distribution

We now propose specific priors for the parameters $(\gamma, \delta_1, \delta_2)$ in (6.9), (6.12), and (6.13) for the case when f in (6.8), and its corresponding subfamilies, is a Student-t distribution with δ degrees of freedom. The shape parameter δ in the Student–t distribution has a clear role since it controls the peakedness and the heaviness of tails of the density function. For instance, the Student–t distribution contains the Cauchy distribution as a particular case for $\delta = 1$, and the normal distribution as a limit case when $\delta \to \infty$. The choice of this sampling model is motivated as follows. Firstly, in Section 6.2.2 we showed that the parameters (δ_1, δ_2) of the TPSH Student-t mainly control the shape of distribution in each direction, while the parameter γ of the TPSC Student-t controls the mass cumulated on each side of the mode. Since the DTP Student-t contains these two models, a similar interpretation for the role of $(\gamma, \delta_1, \delta_2)$ can be provided for this family. Then, this interpretability of the parameters facilitates the choice of hyperparameters of the Bayesian models proposed in Section 6.3. Secondly, since the Student-t distribution is a scale mixture of normals, we can apply the results presented in Section 6.3 on the existence of the posterior distribution under the use of these prior structures. In addition, for DTP and TPSC models we employ the parameterisation in Mudholkar and Hutson (2000) $\{a(\gamma), b(\gamma)\} = \{1 - \gamma, 1 + \gamma\}$, $\gamma \in (-1, 1)$.

Priors for TPSC models

For $p(\gamma)$ and $p(\delta)$ in (6.9) we assume that

$$\begin{aligned} \gamma &\sim \text{Unif}(-1, 1), \\ \delta &\sim \text{TBeta}_2(d, 2, 1), \end{aligned} \tag{6.14}$$

where $\text{TBeta}_2(d, 2, 1)$ is a Beta prime distribution with scale parameter d and shape parameters $(2, 1)$, truncated on $(0.1, \infty)$. The $\text{TBeta}_2(d, 2, 1)$ density can be written, up to a proportionality constant, as

$$p(\delta) \propto \frac{\delta}{(d + \delta)^3} I(\delta > 0.1). \tag{6.15}$$

The hyperparameter d controls the mode $d/2$ and the median $1 + \sqrt{2}d$. Since the parameter δ controls the heaviness of tails of the Student-t distribution, one can include prior beliefs on this parameter by placing the mode of (6.15) on a reasonable value. The untruncated version of this prior was proposed in Juárez and Steel (2010) as a distribution that resembles the tails of the Jeffreys prior of this parameter in the symmetric case. The truncation here is introduced in order to satisfy the conditions in (6.11) given that some of the data sets used in the examples in Section 6.5 contain repeated observations. These conditions are satisfied by the particular choice in (6.14) whenever $k < \dfrac{n+10}{11}$, which represents approximately the 10% of the sample size for samples containing more than 30 observations. The truncation does not represent a strong limitation since it only discards models with very heavy tails which may not be of interest in practice. It is worth pointing out that other levels of truncation can be imposed on the prior (6.15), as long as condition (6.11) is satisfied. The choice of the truncation at 0.1 was simply made for illustrative purposes. Under this parameterisation we have that $AG(\gamma) = -\gamma$, so the uniform prior on

γ also implies a uniform prior on the AG measure of skewness.

Priors for TPSH models

For $p(\delta_1, \delta_2)$ in (6.12) we adopt a hierarchical prior structure

$$\begin{aligned} (\delta_1, \delta_2)|\rho &\sim GC(P_1, P_2; \rho), \\ \rho &\sim P_\rho. \end{aligned} \tag{6.16}$$

where P_1 and P_2 are proper priors, GC denotes a Gaussian Copula distribution with marginals P_1 and P_2, correlation parameter $\rho \in (-1, 1)$, and P_ρ is a proper prior on ρ. This prior structure allows modelling dependencies between δ_1 and δ_2 through the parameter ρ. In the examples presented in 6.5 we assume that P_1 and P_2 are Beta prime distributions truncated on $(0.1, \infty)$ with scale parameters d_1 and d_2, respectively.

For a bivariate Gaussian copula the Spearman's measure of association, $r_\rho \in (-1, 1)$, can be calculated in closed form as (Carta and Steel, 2012)

$$r_\rho = \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right).$$

Using this result, we assume a uniform prior $r_\rho \sim U(-1, 1)$, which leads to the following prior density on ρ

$$p(\rho) \propto \frac{1}{1 - (\rho/2)^2}. \tag{6.17}$$

Priors for DTP models

Finally, for $p(\gamma, \delta_1, \delta_2)$ in (6.13) we additionally assume the product structure $p(\gamma, \delta_1, \delta_2) = p(\gamma)p(\delta_1, \delta_2)$. For the choice of each term we use a combination of the priors proposed for TPSC and TPSH models, leading to the following prior structure

$$\begin{aligned} (\delta_1, \delta_2)|\rho &\sim GC(P_1, P_2; \rho), \\ \rho &\sim P_\rho, \\ \gamma &\sim \text{Unif}(-1, 1). \end{aligned} \tag{6.18}$$

where P_1 and P_2 are Beta prime distributions truncated on $(0.1, \infty)$ with scale parameters

d_1 and d_2, respectively, and ρ is distributed as in (6.17).

6.5 Examples

Now we present four examples with real data to illustrate the use of the Bayesian models proposed for DTP, TPSC and TPSH Student-t distributions in Section 6.4. The choice for the values of the hyperparameters (d, d_1, d_2) is discussed in each example. Simulations of the posterior distributions are obtained using the t-walk algorithm (Christen and Fox, 2010) using a burn-in period of $50,000$ iterations and a thinning of 100 iterations. In each example we construct smoothed marginal posterior distributions of δ, δ_1, and δ_2 using a kernel density estimator and the corresponding posterior samples. These, and the corresponding marginal prior distributions are displayed on the same graph in order to assess the impact of the prior information. Model comparison is conducted via Bayes factors (BF) which are obtained using an importance sampling technique (Chopin and Robert, 2010). In order to assess the Monte Carlo variability of this approximation, boxplots obtained with 100 replications of the BF are presented in each example. The predictive density is used in all the examples as a simple visual tool to evaluate the fit of the models.

6.5.1 Fibre Glass Strength (Similar Tails/Different Cumulated Mass on Each Side of the Mode)

In our first example we analyse the popular data set presented in Smith and Naylor (1987) about the breaking strength of $n = 63$ glass fibres of length 1.5 cm. Since the use of heavy tailed distributions have been suggested for modelling this kind of data, we employ the hyperparameter values $d = d_1 = d_2 = 6$ for the priors in (6.14), (6.16), and (6.18). These hyperparameters produce marginal priors $p(\delta)$ with mode at 3 and median 5.2, then favoring values of δ associated with heavy tails. Figure 6.4 shows the corresponding marginal density. From Figure 6.5 we can observe that the impact of the prior information is significant on the parameter δ_2 of the TPSH model since the shape of the posterior resembles that of the prior. On the other hand, the prior information seems to be moderately influential in the remaining cases. This suggests that the profile likelihood surface of these parameters may be flat and that a larger sample is necessary in order to learn about the tail behaviour in each direction.

For completeness, we also compare these models with the skew-t distributions in Azzalini and Capitanio (2003) and Jones and Faddy (2003) using AIC and BIC. The skew-t density proposed by Azzalini and Capitanio (2003) is given by

$$s_{AC}(x;\mu,\sigma,\lambda,\delta) = \frac{2}{\sigma} f\left(\frac{x-\mu}{\sigma};\delta\right) F\left(\lambda\frac{x-\mu}{\sigma};\delta\right),$$

where $\lambda \in \mathbb{R}$, $f(\cdot;\delta)$ and $F(\cdot;\delta)$ are the Student-t PDF and CDF, respectively. The parameter δ mainly controls the tails of the distribution while the parameter λ is interpreted as a skewness parameter since s_{AC} converges to a right/left-half Student-t distribution as $\lambda \to \pm\infty$.

The skew-t density from Jones and Faddy (2003) is defined as

$$s_{JF}(x;\mu,\sigma,a,b) = C_{a,b}^{-1}\left[1+\frac{t}{\sqrt{a+b+t^2}}\right]^{a+1/2}\left[1-\frac{t}{\sqrt{a+b+t^2}}\right]^{b+1/2},$$

where $a,b>0$, $C_{a,b} = 2^{a+b-1}\,\mathrm{Beta}(a,b)\sqrt{a+b}$, and $t = \dfrac{x-\mu}{\sigma}$. The parameters (a,b) control the tails and skewness jointly. The density s_{JF} is asymmetric if and only if $a \neq b$, which implies that the density is skewed only when the tail behaviour if different in each direction (Jones and Faddy, 2003).

Table 6.2 shows the corresponding maximum likelihood estimators (MLE) of the five skew-t models. We can notice that the MLE of DTP model suggest a similar tail behaviour in each direction and that the skewness is originated by a difference in the mass cumulated on each side of the mode. This is also supported by AIC and BIC criteria which favor the TPSC model. The BF shown in Figure 6.6b slightly favour TPSC model as well. Figure 6.6a shows the predictive densities of the DTP, TPSC and TPSH models.

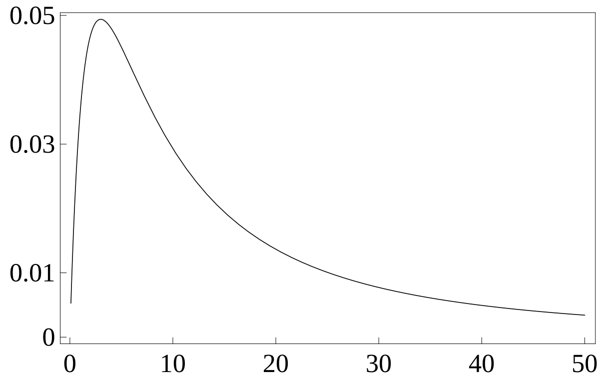


Figure 6.4: Prior on δ for $d = 6$.

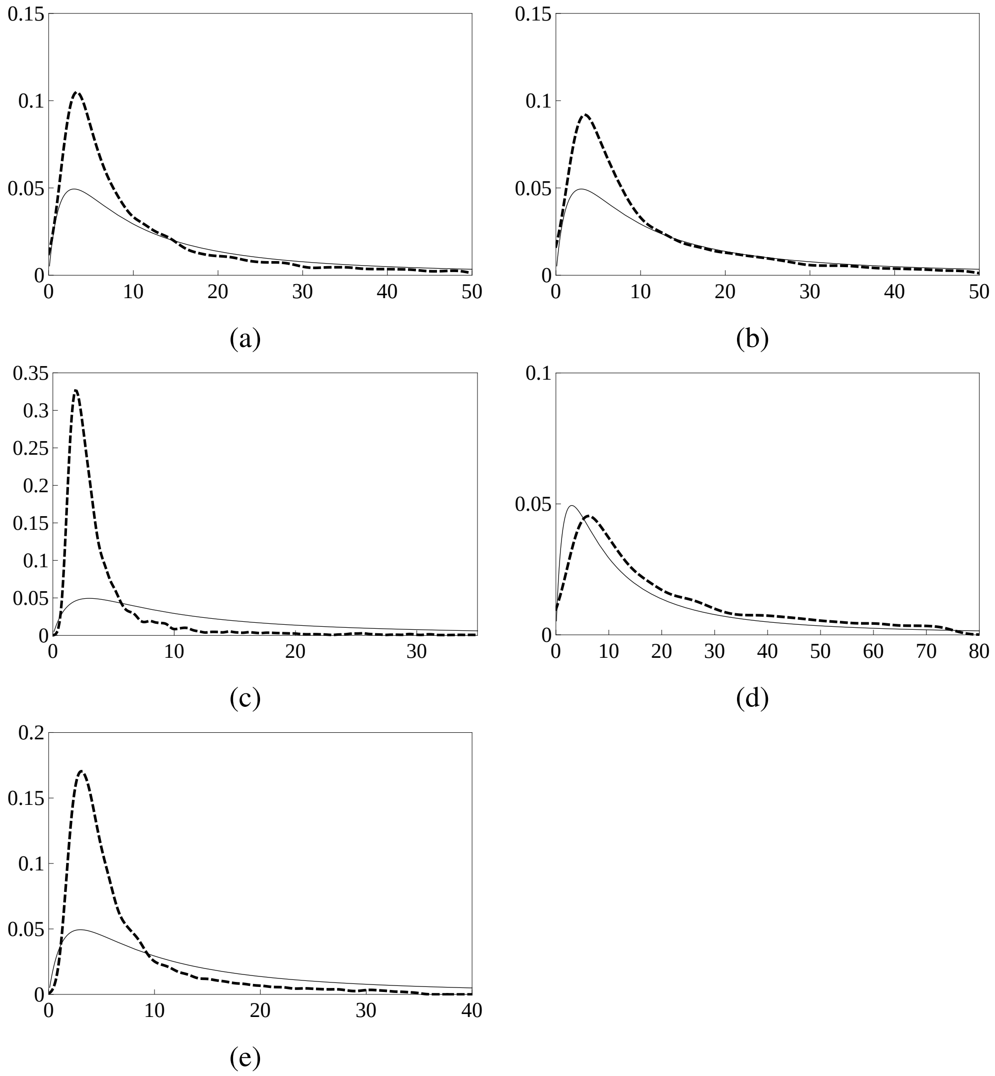


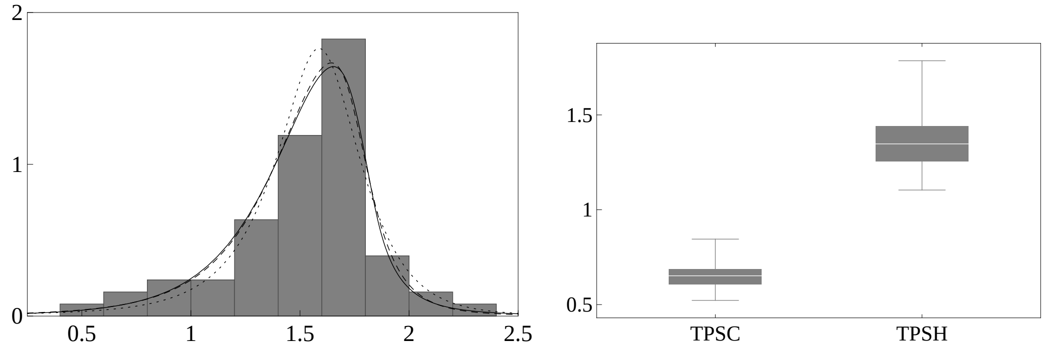
Figure 6.5: Fibre glass strength data. Marginal smoothed posterior (dashed line) and prior distributions (continuous line) for: (a) δ_1 DTP model; (b) δ_2 DTP model; (c) δ_1 TPSH model; (d) δ_2 TPSH model; (e) δ TPSC model.

Model	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\delta}_1$	$\hat{\delta}_2$
DTP	1.66	0.2	0.35	2.72	2.92
TPSC	1.66	0.2	0.35	2.80	–
TPSH	1.58	0.19	–	1.57	7.44
s_{JF}	1.69	0.18	–	($\hat{a}$) 1.11	($\hat{b}$) 2.08
s_{AC}	1.67	0.19	($\hat{\lambda}$) -0.60	2.05	–

Table 6.2: Fibre glass data: Maximum likelihood estimates.

Model	AIC	BIC
DTP	33.6	44.3
TPSC	31.6	40.2
TPSH	33.7	42.3
s_{JF}	31.9	40.4
s_{AC}	31.9	40.4

Table 6.3: Fibre glass data: AIC and BIC criteria.



(a) (b)

Figure 6.6: Fibre glass strength data: (a) Predictive densities: DTP (continuous line); TPSC (dashed line); TPSH (dotted line); (b) Bayes factors: DTP vs. TPSC and DTP vs. TPSH.

6.5.2 Exchange Rates EUR/NOK (Different Tails/Similar Cumulated Mass on Each Side of the Mode)

In this example, we study the log-returns of the exchange rates series EUR/NOK from 01/01/1999 to 01/01/2004 which consist of $n = 1647$ observations. Aas and Haff (2006) analysed these data using a subclass of generalised hyperbolic distributions, which they denoted as GH skew-t, with density given by

$$s_{AA}(x;\mu,\beta,\delta,\nu) = \frac{2^{\frac{1-\nu}{2}}\delta^{\nu}\,|\beta|^{\frac{\nu+1}{2}}\,e^{\beta(x-\mu)}K_{\frac{\nu+1}{2}}\left(\sqrt{\beta^2\left(\delta^2+(x-\mu)^2\right)}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)\left(\delta^2+(x-\mu)^2\right)^{\frac{\nu+1}{4}}},$$

where $\beta \in \mathbb{R}$, $\delta, \nu > 0$ and $K_{\frac{\nu+1}{2}}(\cdot)$ is the modified Bessel function of the third kind of order $\frac{\nu+1}{2}$. The density s_{AA} is asymmetric for $\beta \neq 0$ and coincides with the Student-t distribution for $\beta = 0$. This family contains heavy tailed distributions since its variance exists only if $\nu > 4$. With the exception of the case $\beta = 0$, this model presents different tail behavior in each direction (Aas and Haff, 2006).

Due to the wide variability of this kind of observations, the use of heavy tailed distributions has been suggested for modelling the corresponding log-returns. For this reason, we employ again the hyperparameter values $d = d_1 = d_2 = 6$ for the priors in (6.14), (6.16), and (6.18). From Figure 6.7 we note that the prior distribution has a negligible influence on the shape of the posterior distributions in all the cases. This was expected given that the sample size is large.

Table 6.4 shows the MLE of the four models considered. The MLE of the DTP model suggest that the asymmetry is originated by a different tail behaviour in each direction. This is also reflected in the values of the AIC and BIC criteria which favor the TPSH model. Figure 6.8b shows the BF which strongly favor the TPSH model.

Model	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\delta}_1$	$\hat{\delta}_2$
DTP	2.17×10^{-7}	2.1×10^{-3}	0.06	4.08	2.53
TPSC	-1.9×10^{-4}	2.1×10^{-3}	-0.02	3.08	–
TPSH	-1.5×10^{-4}	2.1×10^{-3}	–	3.60	2.79
s_{AA}	-2.4×10^{-4}	–	($\hat{\beta}$) 18.03	($\hat{\nu}$) 3.11	($\hat{\delta}$) 3.7×10^{-3}

Table 6.4: EUR/NOK exchange rates data: Maximum likelihood estimates.

Model	AIC	BIC
DTP	-14466.9	-14439.9
TPSC	-14464.1	-14442.5
TPSH	-14467.2	-14445.5
s_{AA}	-14466.1	-14444.5

Table 6.5: EUR/NOK exchange rates data: AIC and BIC criteria.

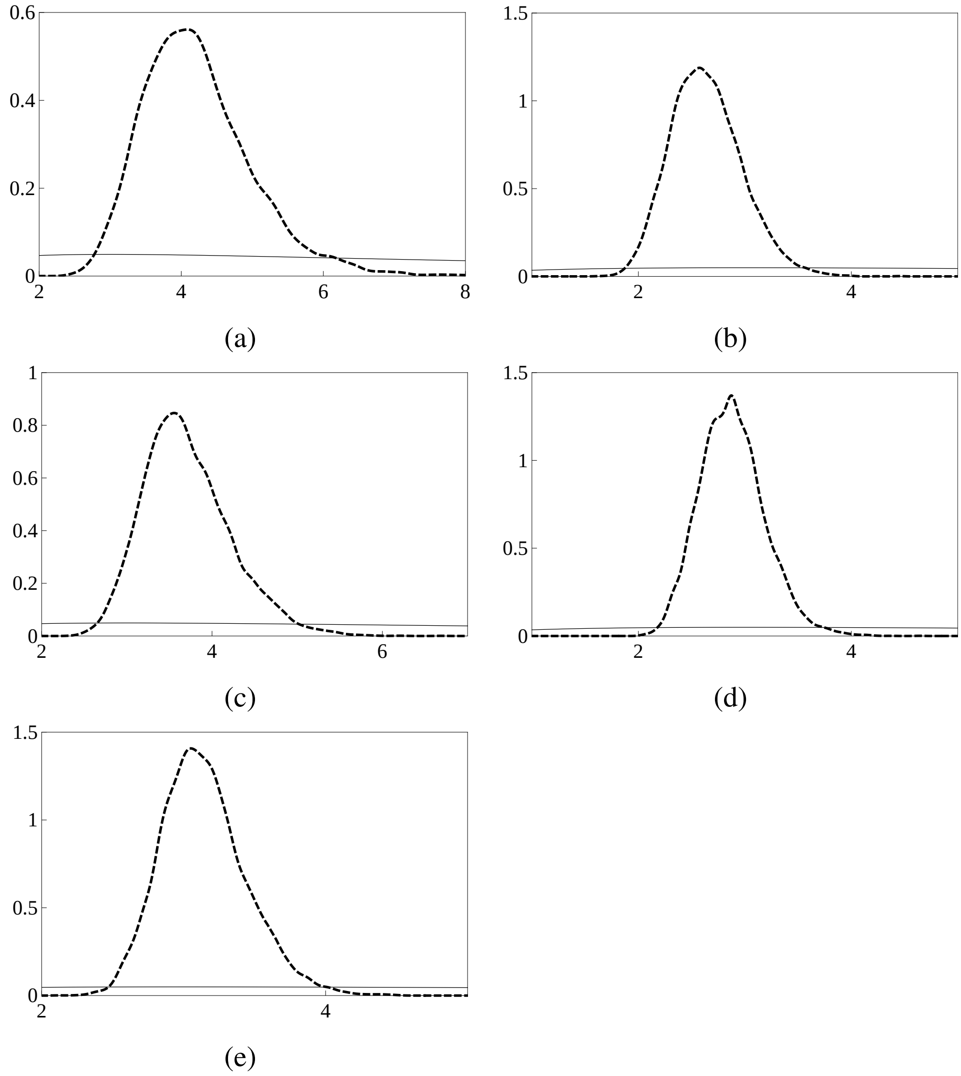


Figure 6.7: EUR/NOK exchange rates. Marginal smoothed posterior (dashed line) and prior distributions (continuous line) for: (a) δ_1 DTP model; (b) δ_2 DTP model; (c) δ_1 TPSH model; (d) δ_2 TPSH model; (e) δ TPSC model.

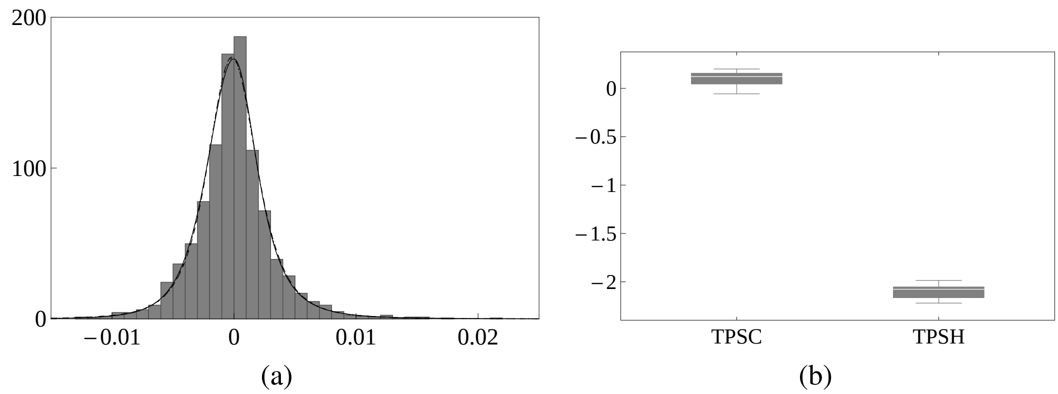


Figure 6.8: EUR/NOK exchange rates: (a) Predictive densities: DTP (continuous line); TPSC (dashed line); TPSH (dotted line); (b) Log–Bayes factors: DTP vs. TPSC and DTP vs. TPSH.

6.5.3 Example 3: Actuarial Application (Similar Tails/Different Cumulated Mass on Each Side of the Mode)

In this application we analyse the claim sizes reported in Berlaint et al. (2004) which can be found in http:// lstat.kuleuven.be/Wiley/. The full data set contains $n = 1823$ observations provided by the reinsurance brokers Aon Re Belgium. This kind of data typically contain extreme observations, whereby the use of the logarithmic transformation is often employed to reduce the effect of these extreme values (Ramirez-Cobo et al., 2010). A quantity of interest in this context is the probability that the claims exceed certain bound κ (Venturini et al., 2008). This information is often used for planning the budget in subsequent years. This emphasises the importance of properly modelling the tails of the distribution.

Since we are interested on the analysis of the logarithm of the observations, we expect that the transformed data present lighter tails than the raw observations. For this reason we employ a common hyperparameter $d = d_1 = d_2 = 20$ in the priors (6.14), (6.16), and (6.18), which produces a density with mode at 10 and median at 29.3. Figure 6.9 shows the corresponding marginal prior density. This prior favors values of δ associated with semi-heavy tailed distributions. From Figure 6.10 we can observe a moderate influence of the prior in the shape of the posterior of δ_1 of the DTP model, while the prior information seems to be less influential in the remaining cases.

We also compare these models with the skew-t distributions in Azzalini and Capitanio (2003) and Jones and Faddy (2003) using AIC and BIC criteria. Table 6.6 shows the MLE of these five models. The MLE of the DTP model suggest that skewness is generated by different mass cumulated on each side of the mode and that the tail behaviour in each direction is similar. The AIC and BIC criteria favor the TPSC model and the skew-t from Azzalini and Capitanio (2003). The BF shown in Figure 6.11 favor the TPSC model as well. Figure 6.11a shows the corresponding predictive densities and suggests a poor fit of the TPSH model which clearly affects the estimation of the right-tail probabilities shown in Figure 6.12.

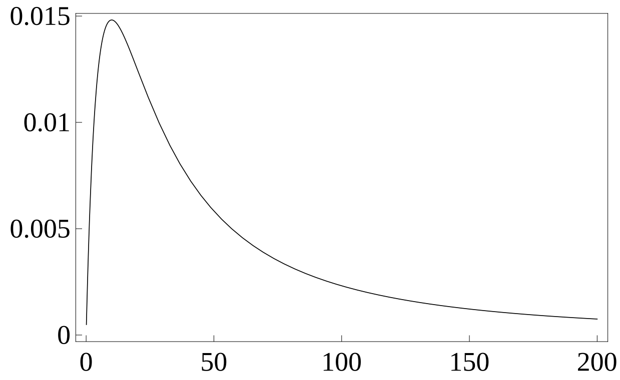


Figure 6.9: Prior on δ for $d = 20$.

Model	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\delta}_1$	$\hat{\delta}_2$
DTP	7.93	1.61	-0.57	16.80	9.87
TPSC	7.90	1.62	-0.59	10.98	–
TPSH	9.13	1.46	–	42945.1	2.50
s_{JF}	1.56	0.02	–	($\hat{a}$) 1560.6	($\hat{b}$) 5.07
s_{AC}	7.17	2.84	($\hat{\lambda}$) 4.90	13.75	–

Table 6.6: Aon data: Maximum likelihood estimates.

Model	AIC	BIC
DTP	7283.1	7310.7
TPSC	7281.6	7303.6
TPSH	7434.4	7456.5
s_{JF}	7302.1	7324.1
s_{AC}	7280.7	7302.7

Table 6.7: Aon data: AIC and BIC criteria.

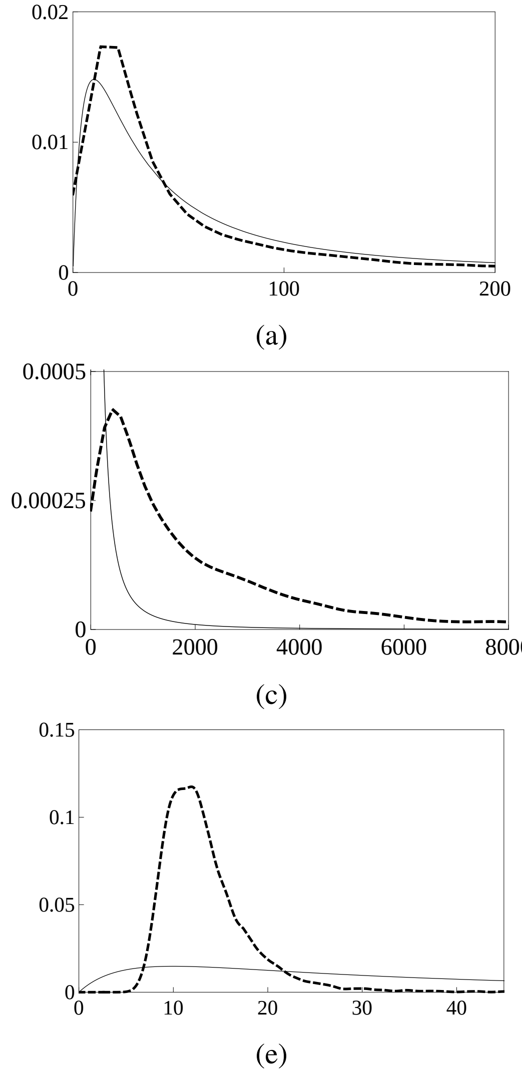


Figure 6.10: Aon data. Marginal smoothed posterior (dashed line) and prior distributions (continuous line) for: (a) δ_1 DTP model; (b) δ_2 DTP model; (c) δ_1 TPSH model; (d) δ_2 TPSH model; (e) δ TPSC model.

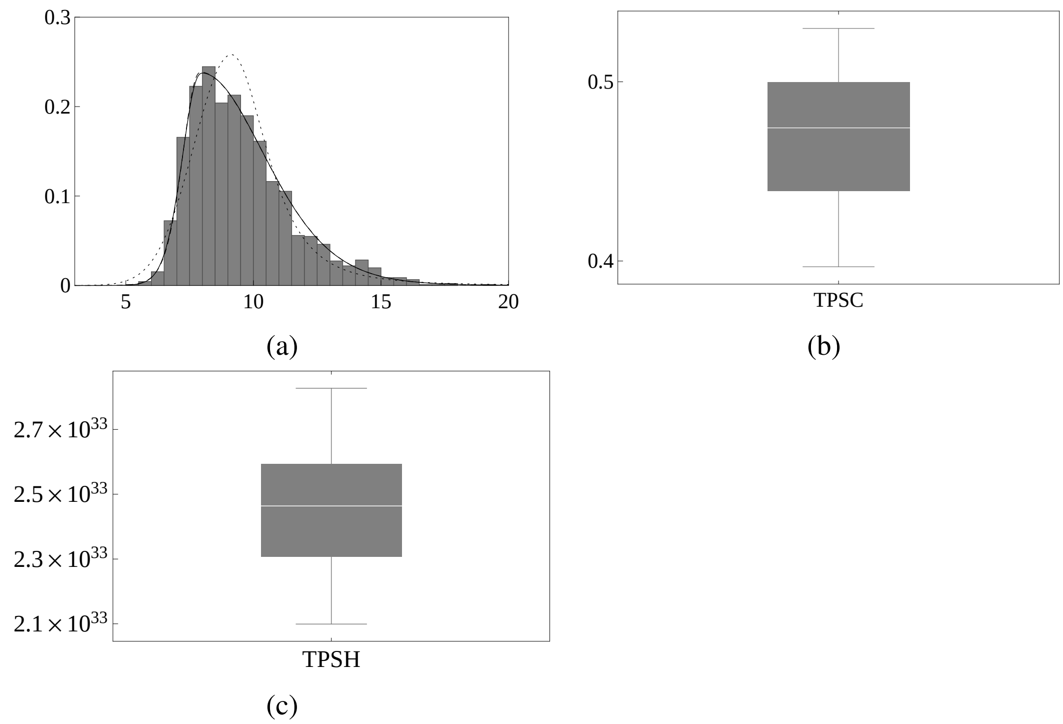


Figure 6.11: Aon data: (a) Predictive densities: DTP (continuous line); TPSC (dashed line); TPSH (dotted line); (b) Bayes factors: DTP vs. TPSC; (c) Bayes factors: DTP vs. TPSH.

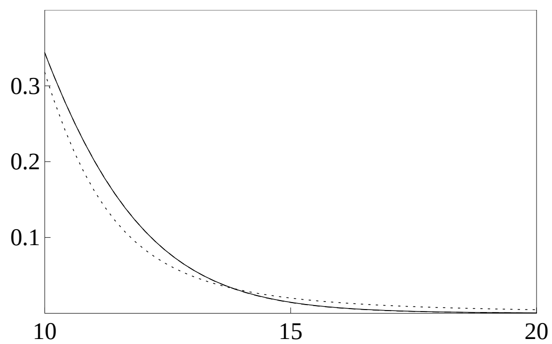


Figure 6.12: Predictive right tail probabilities Aon data: DTP (continuous line); TPSC (dashed line); TPSH (dotted line).

6.5.4 Example 4: Biometric Measurements (Different Tails/Different Cumulated Mass on Each Side of the Mode)

In this example we study the variable "waist girth" of the data set published in Heinz et al. (2003). This data set contains 260 observations measured on physically active women.

In order to come up with hyperparameters for the priors (6.14), (6.16), and (6.18) we consider the following remark. Heinz et al. (2003) noted that "in a well-nourished group the lower limit of waist girth will not fall more than a few centimeters below what can be expected from body build, but the upper limit of waist girth is determined by fatness in addition to body build". Thus, one can expect to observe a normal-like left-shoulder of the distribution of this variable while the right tail is expected to be heavier. For this reason we use the hyperparameter values: $d = 10$ for (6.14), and $(d_1, d_2) = (100, 10)$ for (6.16) and (6.18). Figure 6.13 illustrates the corresponding marginal densities. Note that $d_2 = 10$ favors values of δ associated with semi-heavy tails. On the other hand, $d_1 = 100$ favors values of δ that produce normal-like distributions. This is in line with the prior beliefs about the left and right tails. Figure 6.14 indicates that the prior distribution is moderately influential on the shape of the posterior of δ_1 of the DTP model, while the prior beliefs seem to be less influential in the remaining cases.

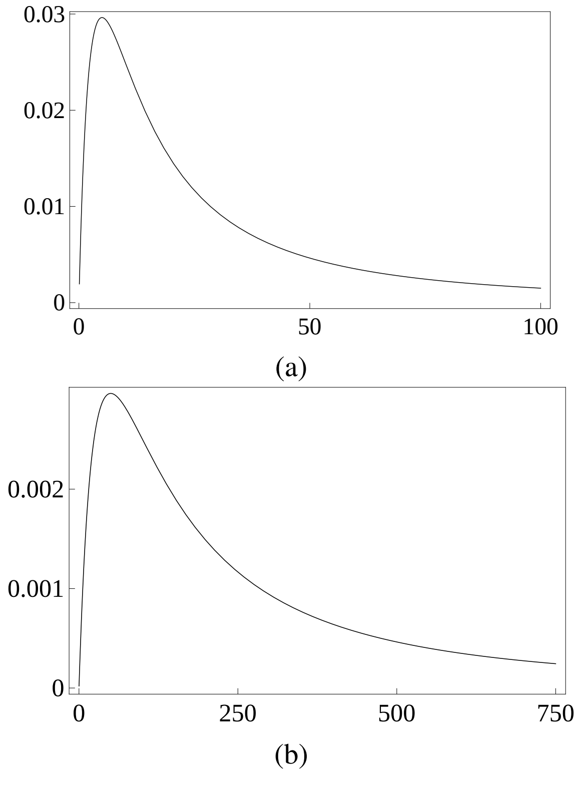


Figure 6.13: Marginal priors on δ for: (a) $d = 10$; (b) $d = 100$.

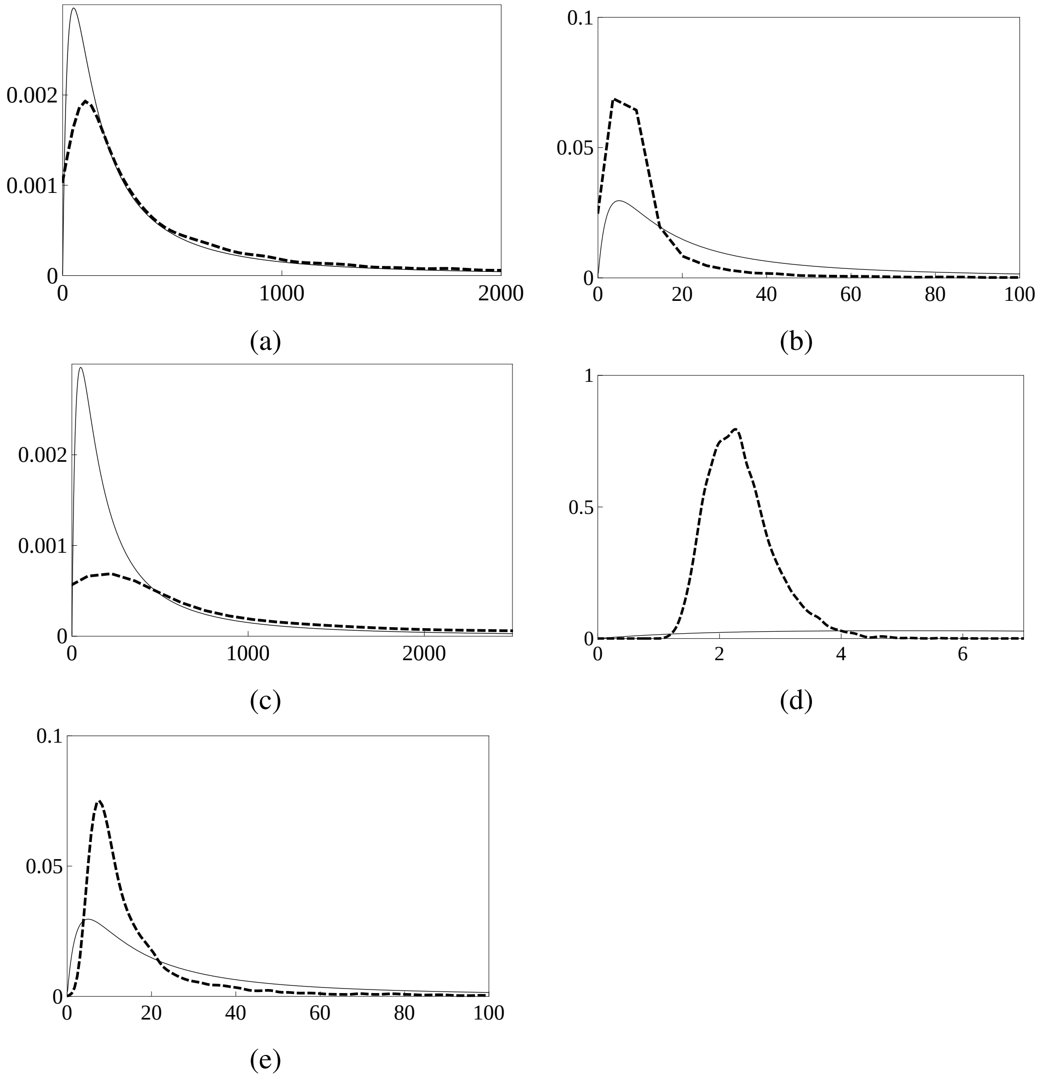


Figure 6.14: Waist girth data. Marginal smoothed posterior (dashed line) and prior distributions (continuous line) for: (a) δ_1 DTP model; (b) δ_2 DTP model; (c) δ_1 TPSH model; (d) δ_2 TPSH model; (e) δ TPSC model.

For completeness, we also compare these models with the skew-t distribution in Azzalini and Capitanio (2003) using AIC and BIC criteria. Table 6.8 shows the corresponding MLE which suggest that skewness comes from both different tail behaviour in each direction and different mass cumulated on each side of the mode. AIC favors the s_{AC} which is followed by the DTP model. BIC, on the other hand, favors the s_{AC}, followed by the TPSC model. This is a consequence of the strong penalty of this criterion on the number of model parameters. Figure 6.15 shows the BF which slightly favor the DTP model. A common point of all these criteria is the inclination for an asymmetric model with different tails in each direction.

Model	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\delta}_1$	$\hat{\delta}_2$
DTP	64.8	5.79	-0.44	41870.9	5.08
TPSC	63.89	5.90	-0.56	7.73	–
TPSH	68.11	5.16	–	57777.8	2.06
s_{AC}	61.40	9.90	($\hat{\lambda}$) 4.00	8.69	–

Table 6.8: Waist girth data: Maximum likelihood estimates.

Model	AIC	BIC
DTP	1738.6	1756.4
TPSC	1738.9	1753.1
TPSH	1748.1	1762.4
s_{AC}	1736.6	1750.9

Table 6.9: Waist girth data: AIC and BIC criteria.

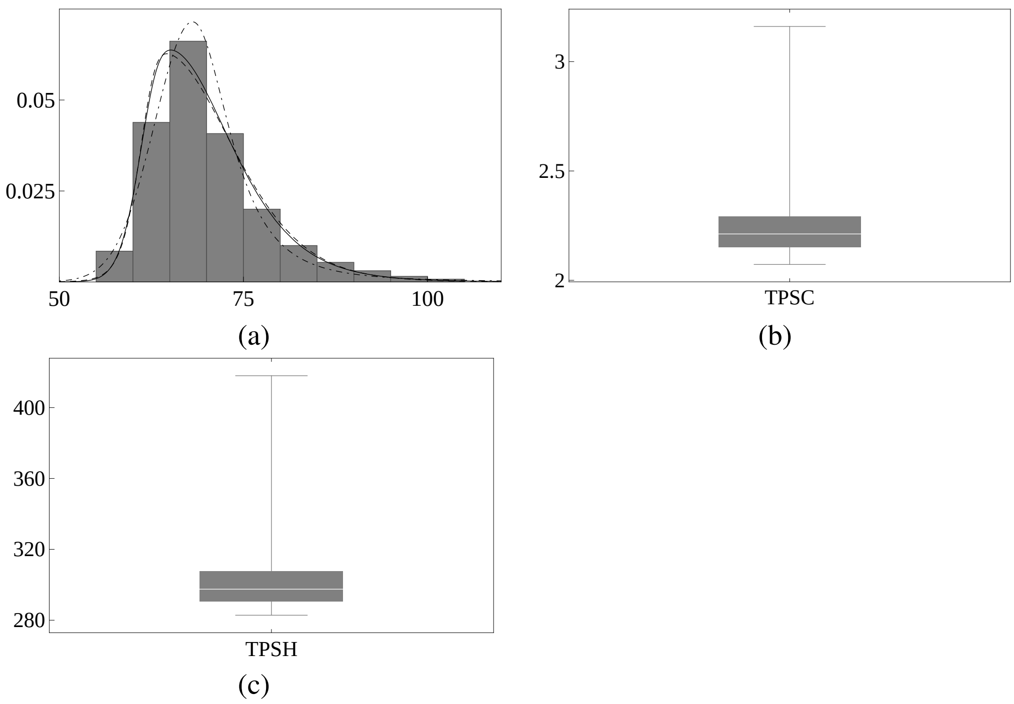


Figure 6.15: Waist girth data: (a) Predictive densities: DTP (continuous line); TPSC (dashed line); TPSH (dotted line); (b) Bayes factors: DTP vs. TPSC; (c) Bayes factors: DTP vs. TPSH.

6.6 Discussion

We have introduced a simple and general class of transformations (DTP) that produces unimodal flexible distributions with parameters that control skewness and shape on each side of the mode. Although some particular cases of DTP models have already been published (Zhu and Galbraith, 2010, 2011), we have formalised this idea and extended it to the family of distributions $\mathcal{F}$. We have also shown that this family contains two subclasses of transformations that can be interpreted as skewing mechanisms. An advantage of the DTP class of transformations is the interpretability of its parameters which, in the Bayesian context, facilitates the translation of prior beliefs into a prior distribution. We proposed a scale-and-location invariant prior structure and showed some guidelines for the choice of the corresponding hyperparameters through examples with real data.

The DTP, TPSC and TPSH models can be used to construct robust models but also, since they capture different sorts of asymmetry, conducting model selection between these models provides more insights on the features governing the behaviour of a data set. We considered Bayes factors as a Bayesian model choice technique but other tools assessing different properties, such as log-predictive scores, might be considered as well.

A different subclass of DTP transformations can be obtained by fixing $\sigma_1 = \sigma$ and $\sigma_2 = \dfrac{f(0;\delta_2)}{f(0;\delta_1)}\sigma$. This sort of transformation produces distributions with different shapes but equal mass cumulated on each side of the mode. This idea is explored in Appendix F. We also show that this transformation can be composed with certain skewing mechanisms to produce a different type of generalised skew-t distribution.

In Chapter 4 we explored the use of Jeffreys priors on TPSC models. Although we partially explore this for TPSH models in Appendix E, further research on the use of Jeffreys priors on TPSH and DTP models is needed. The study of the multivariate extensions proposed in Section 6.2.4 as well as appropriate Bayesian models for these also point out lines for future research.

Chapter 7

Conclusion

7.1 Summary and Conclusions

We have explored inferential and distributional aspects of some classes of flexible distributions obtained by adding parameters to symmetric models.

In Chapter 2, we showed that two parametric transformations that have recently been recommended as skewing mechanisms produce distributions that cannot generally accommodate substantial skewness. Using these results we can identify some desirable properties of a flexible distribution:

(i) **Unimodality**. The reason for this requirement is that the use of finite mixtures is preferred for modelling multimodal data, since multimodality is typically an indicator of the presence of several populations in the sample. This property has also been suggested by Jones and Pewsey (2009).

(ii) **Flexibility with respect to an interpretable measure of kurtosis or skewness**. It is important to employ interpretable measures that allow the user to identify those models that can accommodate moderate or substantial kurtosis/skewness. The use of several measures might be necessary in some cases since different measures capture different features of the model.

(iii) **CDF and PDF in closed form**. Although this is not a strict requirement, this property facilitates the implementation of a model. This requirement can be relaxed to include those distributions involving special functions that are easy to evaluate.

(iv) **Interpretability of the parameters**. This is a useful property from a Bayesian perspective since this facilitates the choice of a prior distribution for the model parameters. From a classical perspective this may also help to interpret the estimates. In addition, this property may also help to avoid adding "redundant" parameters that control similar features, following a principle of parsimony.

These simple properties can be useful as a tool to reduce the long list of "flexible distributions" available in the literature nowadays, and to identify the appropriate models in applications of interest. The distributions obtained by the two-piece transformation satisfy these requirements, which has partially motivated our interest in this kind of models in Chapters 3–6.

Benchmark and noninformative priors are important for Bayesian practitioners in cases where little prior information is available. These priors are also used for comparison purposes since they typically produce inference that is inspired by the shape of the likelihood function rather than by the prior distribution (Robert, 2007). This has motivated the study presented in Chapters 3-5. In Chapter 3 we introduced a benchmark prior for the two–piece Laplace distribution whose structure is inspired by the Jeffreys prior in the symmetric case together with the interpretation of the parameters of the asymmetric model. This Bayesian model was used to model censored observations in the context of microbiology. We then proceeded to study the Jeffreys-rule and the independence Jeffreys priors for two-piece models in a more general framework. We have shown that the Jeffreys-rule prior produces improper posteriors for two-piece scale mixtures of normals sampling models while the existence of the posterior for the choice of the independence Jeffreys prior depends on the parameterisation. Inspired by the structure of the latter, we proposed a benchmark prior structure that produces proper posteriors for two-piece scale mixtures of normals sampling models. We showed through a simulation study that this prior produces posteriors with nice frequentist properties. In Chapter 5 we explored a bivariate extension of two-piece and skew-symmetric models using copulas. We also proposed benchmark priors in this context and presented an application on stress-strength models.

Finally, in Chapter 6 we investigated an extension of the two-piece transformation defined on the family of unimodal location-scale distributions that contain a shape parameter. The resulting distributions contain five interpretable parameters that induce quite a bit of flexibility. We also presented an analysis of certain subfamilies of this transformation and proposed benchmark priors for the corresponding parameters. Conducting model selection between this class of models and the described subclasses does not only provide information about which distribution fits the data better but, since each class captures different features, it also provides more insight about the features governing the phenomenon of interest.

7.2 Future Work

There are some simple modifications to the methods or choices we employed in the thesis that may give place to future research. These, as well as some additional extensions are summarised below.

- The paradigm described in Chapter 2 is applicable to other distributions. In fact, we recommend assessing the flexibility of the asymmetric models of interest using several interpretable measures of kurtosis and/or skewness.

- In Chapter 3 we employed a skew-Laplace for modelling the bacterial size. A natural extension consists of using the skew-exponential power model to obtain inferences about the power parameter (recall that the Laplace distribution is a particular case of the exponential power distribution for $\delta = 1$, where δ is the power parameter). In the context of flow cytometry, data sets are typically large (1000+) and therefore they may contain reliable information about this parameter. For this purpose it would be necessary to propose appropriate priors. It would also be desirable to obtain replications of the experiment in order to evaluate the sampling variability.

- The joint models proposed in Chapter 5 were based on the use of a Gaussian copula. There is a large menu of other bivariate copulas that could be used instead. This, together with the study of appropriate Bayesian models, points out another research line.

- Chapter 6 suggests a couple of extensions. For instance, the study of other particular choices for $f \in \mathcal{F}$ in (6.1) may deserve further study. We believe that a potentially interesting candidate is the symmetric sinh-arcsinh distribution, since this model contains the normal distribution as a particular case as well as models with heavier and lighter tails. The study of distributional and inferential properties of the multivariate extensions proposed in Section 6.2.4 will be considered in the near feature. Finally, a more general extension consists of the application of these multivariate distributions in the context of regression models. For instance, the use of flexible distributions for the distribution of the errors of linear mixed models has recently gained interest (Lachos et al., 2010).

Appendix A

Proofs for Chapter 3

The outline of the proofs of Theorems 2 - 5 is as follows. First, in each case an upper bound of the likelihood function is given. This upper bound depends only on the scale parameter σ and the skewness parameter γ. Using this result it is shown that the integral over the parameter space of the product of the likelihood function and the prior is finite, which implies the properness of the posterior distribution. Throughout, the order statistics of the observations will be denoted by $y^*_{(1)} < y^*_{(2)} \cdots < y^*_{(k)}$.

Proof of Theorem 2

First of all, given that F_X is a CDF, it follows that the contribution of each order statistic $y^*_{(j)}$, $j = 1, \ldots, k$, to the likelihood function is bounded by 1. This is

$$F_X\left(y^*_{(j)} + d; \mu, \sigma, \gamma\right) - F_X\left(y^*_{(j)} - d; \mu, \sigma, \gamma\right) \leq 1.$$

Using this upper bound together with the Mean Value Theorem, it follows that if $0 \leq \mu < y^*_{(2)}$, then

$$\mathcal{L}(y; \mu, \sigma, \gamma) \leq F_X\left(y^*_{(k)} + d; \mu, \sigma, \gamma\right) - F_X\left(y^*_{(k)} - d; \mu, \sigma, \gamma\right) = 2df_X(\zeta_k; \mu, \sigma, \gamma),$$

where $\zeta_k \in \left(y^*_{(k)} - d, y^*_{(k)} + d\right)$. Now, using that this upper bound is an increasing function of μ we can obtain an upper bound that does not depend of μ as indicated below

$$\mathcal{L}(y; \mu, \sigma, \gamma) < 2df_X\left(y^*_{(k)} - d; \mu, \sigma, \gamma\right) < 2df_X\left(y^*_{(k)} - d; y^*_{(2)}, \sigma, \gamma\right).$$

Applying an analogous argument for the range $y^*_{(2)} \leq \mu \leq M$ we obtain a similar upper bound

$$\mathcal{L}(y;\mu,\sigma,\gamma) \;\;\leq\;\; F_X\left(y^*_{(1)}+d;\mu,\sigma,\gamma\right) - F_X\left(y^*_{(1)}-d;\mu,\sigma,\gamma\right) = 2df_X(\zeta_1;\mu,\sigma,\gamma),$$

where $\zeta_1 \in \left(y^*_{(1)}-d, y^*_{(1)}+d\right)$, then

$$\mathcal{L}(y;\mu,\sigma,\gamma) < 2df_X\left(y^*_{(1)}+d;\mu,\sigma,\gamma\right) < 2df_X\left(y^*_{(1)}+d;y^*_{(2)},\sigma,\gamma\right).$$

Therefore we can obtain the following upper bound, for some finite and positive constant C, for the normalising constant of the posterior

$$\begin{aligned}
& \int_0^M \int_0^\infty \int_0^\infty \mathcal{L}(y;\mu,\sigma,\gamma)\pi(\mu,\sigma,\gamma)\, d\sigma d\gamma d\mu \\
= \;\; & \int_0^{y^*_{(2)}} \int_0^\infty \int_0^\infty \mathcal{L}(y;\mu,\sigma,\gamma)\pi(\mu,\sigma,\gamma)\, d\sigma d\gamma d\mu \\
+ \;\; & \int_{y^*_{(2)}}^M \int_0^\infty \int_0^\infty \mathcal{L}(y;\mu,\sigma,\gamma)\pi(\mu,\sigma,\gamma)\, d\sigma d\gamma d\mu \\
\leq \;\; & 2d \int_0^{y^*_{(2)}} \int_0^\infty \int_0^\infty \frac{1}{\sigma^3}\frac{\gamma^2}{(1+\gamma^2)^3} \exp\left(-\frac{y^*_{(k)}-y^*_{(2)}-d}{\sigma\gamma}\right) d\sigma d\gamma d\mu \\
+ \;\; & 2d \int_{y^*_{(2)}}^M \int_0^\infty \int_0^\infty \frac{1}{\sigma^3}\frac{\gamma^2}{(1+\gamma^2)^3} \exp\left(-\gamma\frac{y^*_{(2)}-y^*_{(1)}-d}{\sigma}\right) d\sigma d\gamma d\mu \\
\leq \;\; & C\left(\frac{y^*_{(2)}}{(y^*_{(k)}-y^*_{(2)}-d)^2} + \frac{M-y^*_{(2)}}{(y^*_{(2)}-y^*_{(1)}-d)^2}\right),
\end{aligned}$$

which is finite provided we have at least three distinct observations (*i.e.* $k \geq 3$). This, in turn, implies the properness of the posterior distribution.

Proof of Theorem 3

First of all, note that for all $K_1 \geq \mu$ and $K_2 \geq K_1 + \epsilon$, $\epsilon > 0$

$$\frac{F_X(K_2+\epsilon;\mu,\sigma,\gamma) - F_X(K_2;\mu,\sigma,\gamma)}{F_X(K_1+\epsilon;\mu,\sigma,\gamma) - F_X(K_1;\mu,\sigma,\gamma)} \;\;=\;\; \exp\left[-\frac{K_2-K_1}{\gamma\sigma}\right],$$

and for all $L_2 \leq \mu - \epsilon$ and $L_1 \leq L_2 - \epsilon$

$$\frac{F_X(L_1+\epsilon;\mu,\sigma,\gamma)-F_X(L_1;\mu,\sigma,\gamma)}{F_X(L_2+\epsilon;\mu,\sigma,\gamma)-F_X(L_2;\mu,\sigma,\gamma)} = \exp\left[-\frac{\gamma(L_2-L_1)}{\sigma}\right],$$

If $y^*_{(1)}-d\leq\mu\leq y^*_{(2)}+d$ and $\epsilon=2d$, then

$$\begin{aligned}
\mathcal{L}(y;\mu,\sigma,\gamma,\theta_1,\theta_2) &\leq \frac{F_X\left(y^*_{(k)}+d;\mu,\sigma,\gamma\right)-F_X\left(y^*_{(k)}-d;\mu,\sigma,\gamma\right)}{F_X\left(y^*_{(k)}+d;\mu,\sigma,\gamma\right)-F_X\left(y^*_{(1)}-d;\mu,\sigma,\gamma\right)}\\
&\leq \frac{F_X\left(y^*_{(k)}+d;\mu,\sigma,\gamma\right)-F_X\left(y^*_{(k)}-d;\mu,\sigma,\gamma\right)}{F_X\left(y^*_{(k-1)}+d;\mu,\sigma,\gamma\right)-F_X\left(y^*_{(k-1)}-d;\mu,\sigma,\gamma\right)}\\
&\leq \exp\left[-\frac{y^*_{(k)}-y^*_{(k-1)}}{\gamma\sigma}\right],
\end{aligned}$$

If $y^*_{(2)}+d<\mu\leq y^*_{(k)}+d$ and $\epsilon=2d$, then

$$\begin{aligned}
\mathcal{L}(y;\mu,\sigma,\gamma,\theta_1,\theta_2) &\leq \frac{F_X\left(y^*_{(1)}+d;\mu,\sigma,\gamma\right)-F_X\left(y^*_{(1)}-d;\mu,\sigma,\gamma\right)}{F_X\left(y^*_{(k)}+d;\mu,\sigma,\gamma\right)-F_X\left(y^*_{(1)}-d;\mu,\sigma,\gamma\right)}\\
&\leq \frac{F_X\left(y^*_{(1)}+d;\mu,\sigma,\gamma\right)-F_X\left(y^*_{(1)}-d;\mu,\sigma,\gamma\right)}{F_X\left(y^*_{(2)}+d;\mu,\sigma,\gamma\right)-F_X\left(y^*_{(2)}-d;\mu,\sigma,\gamma\right)}\\
&\leq \exp\left[-\frac{\gamma(y^*_{(2)}-y^*_{(1)})}{\sigma}\right].
\end{aligned}$$

We can then write, for some finite positive C,

$$\int_0^{y^*_{(1)}-d}\int_{y^*_{(k)}+d}^{M}\int_{y^*_{(1)}-d}^{y^*_{(k)}+d}\int_0^\infty\int_0^\infty \mathcal{L}(y;\mu,\sigma,\gamma,\theta_1,\theta_2)\pi(\mu,\sigma,\gamma,\theta_1,\theta_2)\,d\sigma d\gamma d\mu d\theta_1 d\theta_2$$

$$
\begin{aligned}
&\leq \; C \int_{y^*_{(1)}-d}^{y^*_{(2)}+d} \int_0^\infty \int_0^\infty \exp\left[-\frac{y^*_{(k)} - y^*_{(k-1)}}{\gamma\sigma} \right] \frac{1}{\sigma^2} \frac{\gamma}{(1+\gamma^2)^2} \, d\sigma d\gamma d\mu \\
&+ \; C \int_{y^*_{(2)}+d}^{y^*_{(k)}+d} \int_0^\infty \int_0^\infty \exp\left[-\frac{\gamma(y^*_{(2)} - y^*_{(1)})}{\sigma} \right] \frac{1}{\sigma^2} \frac{\gamma}{(1+\gamma^2)^2} \, d\sigma d\gamma d\mu \\
&\propto \; \frac{y^*_{(2)} - y^*_{(1)} + 2d}{y^*_{(k)} - y^*_{(k-1)}} + \frac{y^*_{(k)} - y^*_{(2)}}{y^*_{(2)} - y^*_{(1)}} < \infty, \text{ provided } k \geq 4.
\end{aligned}
$$

The properness of the posterior distribution follows.

Proof of Theorem 4

The proof is analogous to the proof of Theorem 3 using the fact that

$$
1 - F_X\left(y^*_{(1)} - d; \mu, \sigma, \gamma\right) \geq F_X\left(y^*_{(k)} + d; \mu, \sigma, \gamma\right) - F_X\left(y^*_{(1)} - d; \mu, \sigma, \gamma\right).
$$

Proof of Theorem 5

The proof is analogous to the proof of Theorem 3 using the fact that

$$
1 - F_X\left(0; \mu, \sigma, \gamma\right) \geq F_X\left(y^*_{(k)} + d; \mu, \sigma, \gamma\right) - F_X\left(y^*_{(1)} - d; \mu, \sigma, \gamma\right).
$$

Proof of Theorem 6

First we will prove that this result is equivalent to the properness of the posterior distribution for $\gamma = 1$ and then we will prove the result for $\gamma = 1$.

Without loss of generality let us assume that $S_1 \cap S_2 = \emptyset$, this assumption is reasonable given that $k \geq 3$. Then writing the Student's t as a scale mixture of normals with mixing parameters $\lambda = (\lambda_1, \dots, \lambda_n)'$ and applying Fubini's theorem we get an upper bound for $\mathbb{P}[y_1 \in S_1, ..., y_n \in S_n]$ which is proportional to

$$
\begin{aligned}
&\int_{1+\epsilon}^\infty \int_{\mathbb{R}^n_+} \int_{S_n \times \ldots \times S_1} \int_0^\infty \int_0^\infty \int_0^M \left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}} \right) \frac{\sigma^{-n}}{(\gamma + 1/\gamma)^n} \\
\times \; &\exp\left[-\frac{1}{2\sigma^2 h(\gamma)^2} \sum_{j=1}^n \lambda_j (y_j - \mu)^2 \right] \frac{\gamma\sigma^{-2}}{(1+\gamma^2)^2} \, d\mu d\sigma d\gamma dy_1 ... dy_n dP_{\lambda|\nu} dP_\nu,
\end{aligned}
$$

where $h(\gamma) = \max\{\gamma, 1/\gamma\}$. Consider the change of variable $\vartheta = h(\gamma)\sigma$ we can rewrite this upper bound as follows

$$\int_0^\infty \frac{h(\gamma)^{n+1}\gamma^{n+1}}{(1+\gamma^2)^{n+2}}\,d\gamma \int_{1+\epsilon}^\infty \int_{\mathbb{R}_+^n} \int_{S_n\times\ldots\times S_1} \int_0^\infty \int_0^M \left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}}\right) \frac{1}{\vartheta^{n+2}}$$
$$\times \quad \exp\left[-\frac{1}{2\vartheta^2}\sum_{j=1}^n \lambda_j (y_j-\mu)^2\right] d\mu d\vartheta dy_1 ... dy_n dP_{\lambda|\nu} dP_\nu. \tag{A.1}$$

The first integral is finite and the second integral is equivalent to the marginal distribution when $\gamma = 1$. Now we will prove the properness of the posterior distribution for $\gamma = 1$. Defining $S^2(\lambda, y) = \sum_{1\le i<j\le n} \lambda_i\lambda_j(y_i - y_j)^2$ and $\gamma = 1$ we have

$$\int_{1+\epsilon}^\infty \int_{\mathbb{R}_+^n} \int_{S_n\times\ldots\times S_1} \int_0^\infty \int_0^M \left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}}\right) \frac{1}{\sigma^{n+2}} \exp\left[-\frac{1}{2\sigma^2}\frac{S^2(\lambda,y)}{\sum_{j=1}^n \lambda_j}\right]$$
$$\times \quad \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^n \lambda_j\left(\mu - \frac{\sum_{j=1}^n \lambda_j y_j}{\sum_{j=1}^n \lambda_j}\right)^2\right] d\mu d\sigma dy_1 ... dy_n dP_{\lambda|\nu} dP_\nu$$
$$\le \quad \int_{1+\epsilon}^\infty \int_{\mathbb{R}_+^n} \int_{S_n\times\ldots\times S_1} \int_0^\infty \int_{-\infty}^\infty \left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}}\right) \frac{1}{\sigma^{n+2}} \exp\left[-\frac{1}{2\sigma^2}\frac{S^2(\lambda,y)}{\sum_{j=1}^n \lambda_j}\right]$$
$$\times \quad \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^n \lambda_j\left(\mu - \frac{\sum_{j=1}^n \lambda_j y_j}{\sum_{j=1}^n \lambda_j}\right)^2\right] d\mu d\sigma dy_1 ... dy_n dP_{\lambda|\nu} dP_\nu$$
$$\propto \quad \int_{1+\epsilon}^\infty \int_{\mathbb{R}_+^n} \int_{S_n\times\ldots\times S_1} \int_0^\infty \left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}}\right) \left(\sum_{j=1}^n \lambda_j\right)^{-\frac{1}{2}} \frac{1}{\sigma^{n+1}}$$
$$\times \quad \exp\left[-\frac{1}{2\sigma^2}\frac{S^2(\lambda,y)}{\sum_{j=1}^n \lambda_j}\right] d\sigma dy_1 ... dy_n dP_{\lambda|\nu} dP_\nu \tag{A.2}$$

$$\propto \int_{1+\epsilon}^\infty \int_{\mathbb{R}_+^n} \int_{S_n\times\ldots\times S_1} \left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}}\right) \left(\sum_{j=1}^n \lambda_j\right)^{\frac{n-1}{2}} S^2(\lambda,y)^{-\frac{n}{2}}\, dy_1 ... dy_n dP_{\lambda|\nu} dP_\nu. \tag{A.3}$$

Using the proof of Theorem 4 in Fernández and Steel (1998b)

$$S^2(\lambda, y) = \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\left(\sum_{j=1}^{n}\lambda_j\right)\eta_2^2 + (\eta_3-\rho, ..., \eta_n-\rho)Q(\eta_3-\rho, ..., \eta_n-\rho)',$$

where $\eta_i = y_1 - y_i$ for $i = 2, ..., n$, $\rho = \lambda_2\eta_2/(\lambda_1+\lambda_2)$ and $Q = (q_{ij})_{i,j=3}^n$ with diagonal elements $q_{ii} = \lambda_i \sum_{j\neq i}\lambda_j$ and off-diagonal elements $q_{ij} = q_{ji} = -\lambda_i\lambda_j$. Defining $\alpha = \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\left(\sum_{j=1}^n \lambda_j\right)\eta_2^2$ we get

$$\begin{aligned}
S^2(\lambda, y)^{-\frac{n}{2}} &= \alpha^{-\frac{n}{2}}\left[1+(\eta_3-\rho, ..., \eta_n-\rho)\frac{Q}{\alpha}(\eta_3-\rho, ..., \eta_n-\rho)'\right]^{-\frac{n}{2}} \\
&\leq \alpha^{-\frac{n}{2}}\left[1+(\eta_3-\rho, ..., \eta_n-\rho)\frac{Q}{\alpha}(\eta_3-\rho, ..., \eta_n-\rho)'\right]^{-\frac{n-1}{2}} \\
&= \alpha^{-\frac{1}{2}}S^2(\lambda, y)^{-\frac{n-1}{2}} \\
&\leq \left(\lambda_1^{-\frac{1}{2}}+\lambda_2^{-\frac{1}{2}}\right)\left(\sum_{j=1}^n\lambda_j\right)^{-\frac{1}{2}}|\eta_2|^{-1}S^2(\lambda, y)^{-\frac{n-1}{2}}.
\end{aligned}$$

Integrating $(\eta_3, ..., \eta_n)'$ over the whole of $\mathbb{R}^{n-2}$ as in Fernández and Steel (1998b) we get the following upper bound

$$\begin{aligned}
&\int_{S_n\times...\times S_1} S^2(\lambda, y)^{-\frac{n}{2}}\, dy_1...dy_n \leq \left(\lambda_1^{-\frac{1}{2}}+\lambda_2^{-\frac{1}{2}}\right)\left(\sum_{j=1}^n\lambda_j\right)^{-\frac{1}{2}}|\eta_2|^{-1} \\
\times\ &\int_{S_n\times...\times S_1} S^2(\lambda, y)^{-\frac{n-1}{2}}\, dy_1...dy_n \leq \left(\prod_{j=1}^n\lambda_j^{-\frac{1}{2}}\right)\left(\sum_{j=1}^n\lambda_j\right)^{-\frac{n-1}{2}} \\
\times\ &\left(\lambda_1^{-\frac{1}{2}}+\lambda_2^{-\frac{1}{2}}\right)\int_{\{y_1\in S_1, y_1-\eta_2\in S_2\}}|\eta_2|^{-2}\, dy_1 d\eta_2. \qquad \text{(A.4)}
\end{aligned}$$

Combining $(A.3)$ and $(A.4)$ we get

$$\int_{1+\epsilon}^{\infty}\int_{\mathbb{R}_+^n}\int_{S_n\times...\times S_1}\left(\prod_{j=1}^n\lambda_j^{\frac{1}{2}}\right)\left(\sum_{j=1}^n\lambda_j\right)^{\frac{n-1}{2}} S^2(\lambda, y)^{-\frac{n}{2}}\, dy_1...dy_n dP_{\lambda|\nu}dP_\nu$$

$$\leq \int_{1+\epsilon}^{\infty} \int_{\mathbb{R}_+^n} \left(\lambda_1^{-\frac{1}{2}} + \lambda_2^{-\frac{1}{2}} \right) dP_{\lambda|\nu} dP_\nu \int_{\{y_1 \in S_1, y_1 - \eta_2 \in S_2\}} |\eta_2|^{-2} \, dy_1 d\eta_2$$
$$\propto \int_{1+\epsilon}^{\infty} \int_{\mathbb{R}_+} \lambda_1^{-\frac{1}{2}} dP_{\lambda_1|\nu} dP_\nu \int_{\{y_1 \in S_1, y_1 - \eta_2 \in S_2\}} |\eta_2|^{-2} \, dy_1 d\eta_2.$$

The third integral is finite since $S_1 \cap S_2 = \emptyset$. Now, considering that $\lambda_j|\nu \sim Ga\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ for $j = 1, ..., n$

$$\int_{\mathbb{R}_+} \lambda_1^{-\frac{1}{2}} dP_{\lambda_1|\nu} = \frac{\sqrt{2}\Gamma\left(\frac{\nu-1}{2}\right)}{\sqrt{\nu}\Gamma\left(\frac{\nu}{2}\right)} \leq \frac{\sqrt{2}\Gamma\left(\frac{\epsilon}{2}\right)}{\sqrt{\epsilon+1}\Gamma\left(\frac{\epsilon+1}{2}\right)}, \text{ given that } \nu \geq 1+\epsilon.$$

Appendix B

Proofs for Chapter 4 and Supplementary Material

Proof of Theorem 7

The first partial derivatives of $\log[s(y|\mu,\sigma,\gamma)]$ are given by

$$\begin{aligned}
\frac{\partial}{\partial\mu}\log[s(y|\mu,\sigma_1,\sigma_2)] &= -\frac{1}{\sigma_1}\frac{f'\left(\frac{y-\mu}{\sigma_1}\right)}{f\left(\frac{y-\mu}{\sigma_1}\right)}I_{(-\infty,\mu)}(y)-\frac{1}{\sigma_2}\frac{f'\left(\frac{y-\mu}{\sigma_2}\right)}{f\left(\frac{y-\mu}{\sigma_2}\right)}I_{[\mu,\infty)}(y),\\
\frac{\partial}{\partial\sigma_1}\log[s(y|\mu,\sigma_1,\sigma_2)] &= -\frac{1}{\sigma_1+\sigma_2}-\frac{y-\mu}{\sigma_1^2}\frac{f'\left(\frac{y-\mu}{\sigma_1}\right)}{f\left(\frac{y-\mu}{\sigma_1}\right)}I_{(-\infty,\mu)}(y),\\
\frac{\partial}{\partial\sigma_2}\log[s(y|\mu,\sigma_1,\sigma_2)] &= -\frac{1}{\sigma_1+\sigma_2}-\frac{y-\mu}{\sigma_2^2}\frac{f'\left(\frac{y-\mu}{\sigma_2}\right)}{f\left(\frac{y-\mu}{\sigma_2}\right)}I_{[\mu,\infty)}(y).
\end{aligned}$$

Then the entries of the Fisher information matrix of (μ,σ_1,σ_2) are given by

$$
\begin{aligned}
I_{11} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\mu}\log[s(y|\mu,\sigma_1,\sigma_2)]\right)^2\right] = \frac{2\alpha_1}{\sigma_1\sigma_2}, \\
I_{22} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\sigma_1}\log[s(y|\mu,\sigma_1,\sigma_2)]\right)^2\right] = \frac{\alpha_2}{\sigma_1(\sigma_1+\sigma_2)} + \frac{\sigma_2}{\sigma_1(\sigma_1+\sigma_2)^2}, \\
I_{33} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\sigma_2}\log[s(y|\mu,\sigma_1,\sigma_2)]\right)^2\right] = \frac{\alpha_2}{\sigma_2(\sigma_1+\sigma_2)} + \frac{\sigma_1}{\sigma_2(\sigma_1+\sigma_2)^2}, \\
I_{12} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\mu}\log[s(y|\mu,\sigma_1,\sigma_2)]\right)\left(\frac{\partial}{\partial\sigma_1}\log[s(y|\mu,\sigma_1,\sigma_2)\right)\right] = -\frac{2\alpha_3}{\sigma_1(\sigma_1+\sigma_2)}, \\
I_{13} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\mu}\log[s(y|\mu,\sigma_1,\sigma_2)]\right)\left(\frac{\partial}{\partial\sigma_2}\log[s(y|\mu,\sigma_1,\sigma_2)\right)\right] = \frac{2\alpha_3}{\sigma_2(\sigma_1+\sigma_2)}, \\
I_{23} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\sigma_1}\log[s(y|\mu,\sigma_1,\sigma_2)]\right)\left(\frac{\partial}{\partial\sigma_2}\log[s(y|\mu,\sigma_1,\sigma_2)\right)\right] = -\frac{1}{(\sigma_1+\sigma_2)^2}.
\end{aligned}
$$

□

Proof of Theorem 8

The determinant of the Fisher information matrix is

$$
|I(\mu,\sigma_1,\sigma_2)| = \frac{2\alpha_2\left(\alpha_1+\alpha_1\alpha_2-2\alpha_3^2\right)}{\sigma_1^2\sigma_2^2(\sigma_1+\sigma_2)^2}.
$$

We will first prove that $\alpha_2 > 0$. From the definition of α_2 it can only be zero if $1 + tf'(t)/f(t) = 0$ whenever $f(t) > 0$. This means that $f(t) = -tf'(t)$ and this only happens if $f(t) = K/t$ for any positive K. The latter, however, is not a probability density function on $\mathbb{R}$. Thus, α_2 can not be zero.

Next, we will prove that $\alpha_1(1+\alpha_2) > 2\alpha_3^2$. Applying the Cauchy-Schwarz inequality we have $\alpha_1(1+\alpha_2) \geq 2\alpha_3^2$. We will show that this is a strict inequality. The condition in Theorem 8 implies that

$$
0 < \int_0^\infty t\left[\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt.
$$

Let

$$
\phi(t) = \left|\frac{f'(t)}{\sqrt{f(t)}}\right| > 0 \text{ a.e. and } \psi(t) = t\left|\frac{f'(t)}{\sqrt{f(t)}}\right| > 0 \text{ a.e.}
$$

Note that $[\beta\phi(t)+\psi(t)]^2 > 0$ a.e. for any $\beta \in \mathbb{R}$, and thus

$$0 < \int_0^\infty [\beta\phi(t)+\psi(t)]^2\, dt = \beta^2 \int_0^\infty \phi^2(t)\, dt + 2\beta \int_0^\infty \phi(t)\psi(t)\, dt + \int_0^\infty \psi^2(t)\, dt.$$

This is a polynomial of degree 2 in β with positive coefficients and no real roots, implying that the discriminant is negative, so that

$$\left[\int_0^\infty t \left[\frac{f'(t)}{f(t)}\right]^2 f(t)\, dt\right]^2 < \left[\int_0^\infty t^2 \left[\frac{f'(t)}{f(t)}\right]^2 f(t)\, dt\right]\left[\int_0^\infty \left[\frac{f'(t)}{f(t)}\right]^2 f(t)\, dt\right].$$

□

Proof of Theorem 9

The first partial derivatives of $\log[s(y|\mu,\sigma,\gamma)]$ are given by

$$\begin{aligned}
\frac{\partial}{\partial\mu}\log[s(y|\mu,\sigma,\gamma)] &= -\frac{1}{\sigma b(\gamma)}\frac{f'\left(\frac{y-\mu}{\sigma b(\gamma)}\right)}{f\left(\frac{y-\mu}{\sigma b(\gamma)}\right)} I_{(-\infty,\mu)}(y) - \frac{1}{\sigma a(\gamma)}\frac{f'\left(\frac{y-\mu}{\sigma a(\gamma)}\right)}{f\left(\frac{y-\mu}{\sigma a(\gamma)}\right)} I_{[\mu,\infty)}(y),\\
\frac{\partial}{\partial\sigma}\log[s(y|\mu,\sigma,\gamma)] &= -\frac{1}{\sigma} - \frac{y-\mu}{\sigma^2 b(\gamma)}\frac{f'\left(\frac{y-\mu}{\sigma b(\gamma)}\right)}{f\left(\frac{y-\mu}{\sigma b(\gamma)}\right)} I_{(-\infty,\mu)}(y) - \frac{y-\mu}{\sigma^2 a(\gamma)}\frac{f'\left(\frac{y-\mu}{\sigma a(\gamma)}\right)}{f\left(\frac{y-\mu}{\sigma a(\gamma)}\right)} I_{[\mu,\infty)}(y),\\
\frac{\partial}{\partial\gamma}\log[s(y|\mu,\sigma,\gamma)] &= -\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)} - \frac{y-\mu}{\sigma}\frac{b'(\gamma)}{b(\gamma)^2}\frac{f'\left(\frac{y-\mu}{\sigma b(\gamma)}\right)}{f\left(\frac{y-\mu}{\sigma b(\gamma)}\right)} I_{(-\infty,\mu)}(y)\\
&- \frac{y-\mu}{\sigma}\frac{a'(\gamma)}{a(\gamma)^2}\frac{f'\left(\frac{y-\mu}{\sigma a(\gamma)}\right)}{f\left(\frac{y-\mu}{\sigma a(\gamma)}\right)} I_{[\mu,\infty)}(y).
\end{aligned}$$

Thus, the entries of the Fisher information matrix of (μ,σ,γ) are

$$
\begin{aligned}
I_{11} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\mu}\log[s(y|\mu,\sigma,\gamma)]\right)^2\right] = \frac{2\alpha_1}{a(\gamma)b(\gamma)\sigma^2},\\
I_{22} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\sigma}\log[s(y|\mu,\sigma,\gamma)]\right)^2\right] = \frac{\alpha_2}{\sigma^2},\\
I_{33} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\gamma}\log[s(y|\mu,\sigma,\gamma)]\right)^2\right] = \frac{\alpha_2+1}{a(\gamma)+b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)}+\frac{a'(\gamma)^2}{a(\gamma)}\right]-\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right]^2,\\
I_{12} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\mu}\log[s(y|\mu,\sigma,\gamma)]\right)\left(\frac{\partial}{\partial\sigma}\log[s(y|\mu,\sigma,\gamma)\right)\right] = 0,\\
I_{13} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\mu}\log[s(y|\mu,\sigma,\gamma)]\right)\left(\frac{\partial}{\partial\gamma}\log[s(y|\mu,\sigma,\gamma)\right)\right]\\
&= \frac{2\alpha_3}{\sigma[a(\gamma)+b(\gamma)]}\left[\frac{a'(\gamma)}{a(\gamma)}-\frac{b'(\gamma)}{b(\gamma)}\right],\\
I_{23} &= \mathbb{E}\left[\left(\frac{\partial}{\partial\sigma}\log[s(y|\mu,\sigma,\gamma)]\right)\left(\frac{\partial}{\partial\gamma}\log[s(y|\mu,\sigma,\gamma)\right)\right]\\
&= \frac{\alpha_2}{\sigma}\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right].
\end{aligned}
$$

□

Proof of Theorem 10

Note that

$$
\frac{d}{d\gamma}AG(\gamma) = 2\frac{a'(\gamma)b(\gamma)-a(\gamma)b'(\gamma)}{[a(\gamma)+b(\gamma)]^2} = 2\frac{a(\gamma)b(\gamma)\lambda(\gamma)}{[a(\gamma)+b(\gamma)]^2},
$$

so that

$$
\frac{dAG(\gamma)}{d\gamma} > 0 \Leftrightarrow \lambda(\gamma) > 0 \text{ and } \frac{dAG(\gamma)}{d\gamma} < 0 \Leftrightarrow \lambda(\gamma) < 0.
$$

□

Proof of Theorem 11

First of all, consider the independence Jeffreys prior (4.6) and the change of variable (4.7), then

$$
\begin{aligned}
\pi_I(\mu,\sigma,\gamma) &\propto \frac{|a'(\gamma)b(\gamma)-a(\gamma)b'(\gamma)|\sqrt{[b(\gamma)+\alpha_2[a(\gamma)+b(\gamma)]][a(\gamma)+\alpha_2[a(\gamma)+b(\gamma)]]}}{\sigma\sqrt{a(\gamma)b(\gamma)}[a(\gamma)+b(\gamma)]^2} \\
&\leq \frac{(\alpha_2+1)|a'(\gamma)b(\gamma)-a(\gamma)b'(\gamma)|}{\sigma\sqrt{a(\gamma)b(\gamma)}[a(\gamma)+b(\gamma)]}.
\end{aligned}
$$

For the particular choice $\{a(\gamma), b(\gamma)\} = \{\gamma, 1/\gamma\}$, the upper bound of $\pi_I(\mu,\sigma,\gamma)$ is proportional to $[\sigma(1+\gamma^2)]^{-1}$. Now, the proof of (i) and (ii) is as follows.

(i) Applying Theorem 1 from Fernández and Steel (1998b) and using this upper bound we can derive the properness of the posterior distribution of (μ,σ,γ). Now, since the mapping $(\mu,\sigma,\gamma) \leftrightarrow (\mu,\sigma_1,\sigma_2)$ is one-to-one, it follows that the posterior distribution of (μ,σ_1,σ_2) is proper.

(ii) The proof follows analogously by applying Theorem 2 from Fernández and Steel (1998b). □

Proof of Theorem 12

Let f be a scale mixture of normals with λ_j the mixing variable associated with y_j and where the λ_js are independent random variables defined on $\mathbb{R}^+$ with distribution P_{λ_j}.

(i) Integrating with respect of μ over a subspace we get a lower bound for the marginal distribution of $(y_1,...,y_n)$ which is proportional to

$$
\begin{aligned}
&\int_{\mathbb{R}_+^n}\int_\Gamma\int_0^\infty\int_{-\infty}^{y_{(1)}}\left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}}\right)\frac{\sigma^{-(n+1)}}{[a(\gamma)+b(\gamma)]^n}\exp\left[-\frac{1}{2\sigma^2 a(\gamma)^2}\sum_{j=1}^n \lambda_j(y_j-\mu)^2\right] \\
\times\quad & \pi(\gamma)\, d\mu d\sigma d\gamma dP_{(\lambda_1,\ldots,\lambda_n)}.
\end{aligned}
$$

Consider the change of variable $\vartheta = \sigma a(\gamma)$. Then we can rewrite the lower bound as follows

$$
\begin{aligned}
&\int_\Gamma\left[\frac{a(\gamma)}{a(\gamma)+b(\gamma)}\right]^n \pi(\gamma)\, d\gamma \int_{\mathbb{R}_+^n}\int_0^\infty\int_{-\infty}^{y_{(1)}}\left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}}\right)\vartheta^{-(n+1)} \\
\times\quad & \exp\left[-\frac{1}{2\vartheta^2}\sum_{j=1}^n \lambda_j(y_j-\mu)^2\right] d\mu d\vartheta dP_{(\lambda_1,\ldots,\lambda_n)},
\end{aligned}
$$

and the result follows.

(ii) We can get an upper bound for the marginal distribution of $(y_1, ..., y_n)$ proportional to

$$
\begin{aligned}
&\int_{\mathbb{R}_n^+} \int_\Gamma \int_0^\infty \int_{-\infty}^\infty \left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}} \right) \frac{\sigma^{-(n+1)}}{[a(\gamma)+b(\gamma)]^n} \exp\left[-\frac{1}{2\sigma^2 h(\gamma)^2} \sum_{j=1}^n \lambda_j (y_j-\mu)^2 \right] \\
\times \quad & \pi(\gamma)\, d\mu d\sigma d\gamma dP_{(\lambda_1,\dots,\lambda_n)},
\end{aligned}
$$

where $h(\gamma) = \max\{a(\gamma), b(\gamma)\}$. Consider the change of variable $\vartheta = \sigma h(\gamma)$ and rewrite the upper bound as follows

$$
\begin{aligned}
&\int_\Gamma \left[\frac{h(\gamma)}{a(\gamma)+b(\gamma)} \right]^n \pi(\gamma)\, d\gamma \int_{\mathbb{R}_n^+} \int_0^\infty \int_{-\infty}^\infty \left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}} \right) \vartheta^{-(n+1)} \\
\times \quad & \exp\left[-\frac{1}{2\vartheta^2} \sum_{j=1}^n \lambda_j (y_j-\mu)^2 \right] d\mu d\vartheta dP_{(\lambda_1,\dots,\lambda_n)}.
\end{aligned}
$$

Fernández and Steel (2000, Th. 1) show that the integral in $\mu, \vartheta, \lambda_1, ..., \lambda_n$ is finite if $n \geq 2$. Then, by Theorem 1 from Fernández and Steel (1998b), the existence of the integral in γ is a sufficient condition for the properness of the posterior distribution of (μ, σ, γ). The result then follows from

$$
\int_\Gamma \left[\frac{h(\gamma)}{a(\gamma)+b(\gamma)} \right]^n \pi(\gamma)\, d\gamma \leq \int_\Gamma \pi(\gamma)\, d\gamma.
$$

(iii) The proof follows analogously by applying Theorem 2 from Fernández and Steel (1998b).

□

Proof of Theorem 13

If f is a scale mixture of normals, then integrating over a subspace with respect to μ we get a lower bound for the marginal distribution of $(y_1, ..., y_n)$ which is proportional to

$$\int_{\mathbb{R}_+^n}\int_\Gamma\int_0^\infty\int_{-\infty}^{y_{(1)}}\left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}}\right)\frac{\sigma^{-(n+2)}}{[a(\gamma)+b(\gamma)]^n}\exp\left[-\frac{1}{2\sigma^2 a(\gamma)^2}\sum_{j=1}^n \lambda_j (y_j-\mu)^2\right]$$
$$\times\frac{|\lambda(\gamma)|}{a(\gamma)+b(\gamma)}\,d\mu d\sigma d\gamma dP_{(\lambda_1,\ldots,\lambda_n)}.$$

Consider the change of variable $\vartheta=\sigma a(\gamma)$. Then we can rewrite this lower bound as follows

$$\int_\Gamma\left[\frac{a(\gamma)}{a(\gamma)+b(\gamma)}\right]^{n+1}|\lambda(\gamma)|\,d\gamma\int_{\mathbb{R}_+^n}\int_0^\infty\int_{-\infty}^{y_{(1)}}\left(\prod_{j=1}^n \lambda_j^{\frac{1}{2}}\right)\vartheta^{-(n+2)}$$
$$\times\exp\left[-\frac{1}{2\vartheta^2}\sum_{j=1}^n \lambda_j (y_j-\mu)^2\right]\,d\mu d\vartheta dP_{(\lambda_1,\ldots,\lambda_n)}.$$

Therefore, the existence of the first integral is a necessary condition for the properness of the posterior distribution of (μ,σ,γ). □

Proof of Theorem 14

The proof of (i) is as follows. If f is normal, defining $h(\gamma)=\max\{a(\gamma),b(\gamma)\}$ we get an upper bound for the marginal distribution of $(y_1,\ldots,y_n)$ which is proportional to

$$\int_{-\infty}^\infty\int_\Gamma\int_0^\infty\frac{\pi_J(\mu,\sigma,\gamma)}{[a(\gamma)+b(\gamma)]^n\sigma^n}\exp\left[-\frac{1}{2\sigma^2 h(\gamma)^2}\sum_{j=1}^n (y_j-\mu)^2\right]d\sigma d\gamma d\mu$$
$$\propto\int_{-\infty}^\infty\left[\sum_{j=1}^n (y_j-\mu)^2\right]^{-\frac{n+1}{2}}d\mu\int_\Gamma\frac{h(\gamma)^{n+1}}{[a(\gamma)+b(\gamma)]^{n+1}}|\lambda(\gamma)|d\gamma.$$

The first integral exists if $n\geq 2$ and at least 2 observations are different. Then the existence of the second integral is a sufficient condition for the existence of the posterior distribution. For the second integral we use that

$$\int_\Gamma\frac{h(\gamma)^{n+1}}{[a(\gamma)+b(\gamma)]^{n+1}}|\lambda(\gamma)|d\gamma\leq\int_\Gamma|\lambda(\gamma)|d\gamma,$$

which is finite by assumption. If f is Laplace, analogously to the normal case we get an upper bound for the marginal distribution of $(y_1,\ldots,y_n)$ which is proportional to

$$\int_{-\infty}^{\infty}\int_{\Gamma}\int_{0}^{\infty}\frac{\pi_J(\mu,\sigma,\gamma)}{[a(\gamma)+b(\gamma)]^n\sigma^n}\exp\left[-\frac{1}{\sigma h(\gamma)}\sum_{j=1}^{n}|y_j-\mu|\right]d\sigma d\gamma d\mu$$
$$\propto\int_{-\infty}^{\infty}\left[\sum_{j=1}^{n}|y_j-\mu|\right]^{-(n+1)}d\mu\int_{\Gamma}\frac{h(\gamma)^{n+1}}{[a(\gamma)+b(\gamma)]^{n+1}}|\lambda(\gamma)|d\gamma,$$

and the same argument leads to the result.

Result (ii) follows immediately from Corollary 6.

For (iii) let us assume, without loss of generality, that $AG(\gamma)$ is an increasing function and $\Gamma=(\underline{\gamma},\overline{\gamma})$. First, note that we can rewrite $AG(\gamma)$ as follows

$$AG(\gamma)=\tanh\left\{\frac{1}{2}\log\left[\frac{a(\gamma)}{b(\gamma)}\right]\right\}.$$

Then

$$\lim_{\gamma\to\overline{\gamma}}AG(\gamma)=1\quad\Leftrightarrow\quad\lim_{\gamma\to\overline{\gamma}}\log\left[\frac{a(\gamma)}{b(\gamma)}\right]=\infty$$
$$\lim_{\gamma\to\underline{\gamma}}AG(\gamma)=-1\quad\Leftrightarrow\quad\lim_{\gamma\to\underline{\gamma}}\log\left[\frac{a(\gamma)}{b(\gamma)}\right]=-\infty,$$

which contradicts the assumption that $\lambda(\gamma)$ is absolutely integrable. The result is analogous if AG is decreasing. □

Proof of Theorem 15

From Theorem 12(ii) and (iii) we know that properness of $\pi(\gamma)$ in (4.23) is sufficient for existence of the posterior. The AG beta prior implies a proper prior for AG when $\alpha_0,\beta_0>0$. From Theorem 10 the condition that $\lambda(\gamma)$ does not change sign is equivalent to AG being a one-to-one transformation of γ. Thus, the induced prior on γ will be proper and the result follows. □

Appendix 2: Simulation Study

In this section we investigate the empirical coverage of the 95% posterior credible intervals, defined by the 2.5th and 97.5th percentiles. We simulate $N=10,000$ datasets of size $n=30,100$ and 1000 from various sampling models where we take f to be a normal distribution throughout, and analyse these data using the corresponding Bayesian model. Model 1 consists of the two-piece model (4.2) and the independence Jeffreys prior (4.6). Model 2 corresponds to (4.9) using $\{a(\gamma), b(\gamma)\}$ of the ϵ-skew model under the independence Jeffreys prior. Model 3 is the logistic AG model of Example 6 for $\gamma \in [-B, B]$ with the Jeffreys prior in (4.22). Model 4 is the ISF model with the modified Jeffreys prior in (4.26), and Model 5 is the ϵ-skew model in combination with the AG beta prior in (4.29). For each of these N datasets, a sample of size $3,000$ was obtained from the posterior distribution using a Markov chain Monte Carlo sampler after a burn-in period of $5,000$ iterations and thinned to every 50th iteration. Finally, the proportion of 95% credible intervals that include the true value of the parameter was calculated. Results are presented in Tables B.1-B.5. For Model 3 we know that the truncation to a finite interval is what makes the posterior well-defined. To investigate how sensitive the results are to the particular value chosen for B, we have experimented with various values.

Sample size	$n = 30$		$n = 100$		$n = 1000$	
Parameters	$\sigma_1 = 2.0$ $\sigma_2 = 0.5$	$\sigma_1 = 0.66$ $\sigma_2 = 1.50$	$\sigma_1 = 2.0$ $\sigma_2 = 0.5$	$\sigma_1 = 0.66$ $\sigma_2 = 1.50$	$\sigma_1 = 2.0$ $\sigma_2 = 0.5$	$\sigma_1 = 0.66$ $\sigma_2 = 1.50$
μ	0.976	0.967	0.971	0.956	0.948	0.953
σ_1	0.961	0.951	0.974	0.958	0.947	0.949
σ_2	0.975	0.971	0.961	0.951	0.948	0.950

Table B.1: Coverage proportions. Mixture model with independence Jeffreys prior (Model 1)

Sample size	$n = 30$		$n = 100$		$n = 1000$	
Parameter	$\gamma = 0.5$	$\gamma = -0.5$	$\gamma = 0.5$	$\gamma = -0.5$	$\gamma = 0.5$	$\gamma = -0.5$
μ	0.971	0.967	0.954	0.955	0.947	0.948
σ	0.959	0.960	0.947	0.945	0.953	0.954
γ	0.971	0.969	0.957	0.957	0.948	0.952

Table B.2: Coverage proportions. ϵ-skew model with independence Jeffreys prior (Model 2)

Sample size	$n = 30$		$n = 100$		$n = 1000$	
Parameter	$\gamma = 0.5$	$\gamma = -0.5$	$\gamma = 0.5$	$\gamma = -0.5$	$\gamma = 0.5$	$\gamma = -0.5$
μ	0.967	0.964	0.949	0.953	0.948	0.949
σ	0.995	0.991	0.952	0.960	0.948	0.947
γ	0.964	0.965	0.949	0.952	0.948	0.947

Table B.3: Coverage proportions. Logistic AG model with Jeffreys prior (Model 3) and $B = 3$

Size	$n = 30$		$n = 100$		$n = 1000$	
Parameter	$\gamma = 0.5$	$\gamma = 1.5$	$\gamma = 0.5$	$\gamma = 1.5$	$\gamma = 0.5$	$\gamma = 1.5$
μ	0.969	0.967	0.963	0.950	0.949	0.946
σ	0.992	0.972	0.965	0.949	0.947	0.949
γ	0.967	0.971	0.967	0.950	0.950	0.948

Table B.4: Coverage proportions: Inverse scale factors model with modified Jeffreys prior (Model 4)

Size	$n = 30$		$n = 100$		$n = 1000$	
Parameter	$\gamma = 0.5$	$\gamma = -0.5$	$\gamma = 0.5$	$\gamma = -0.5$	$\gamma = 0.5$	$\gamma = -0.5$
μ	0.968	0.967	0.960	0.959	0.947	0.951
σ	0.994	0.993	0.968	0.970	0.947	0.951
γ	0.968	0.969	0.964	0.964	0.948	0.950

Table B.5: Coverage proportions: ϵ-skew model with AG beta prior (Model 5).

All models lead to coverage probabilities above the nominal level for samples of size $n = 30$, especially in the case of σ for Models 3-5. Once we increase the sample size to $n = 100$, the coverage is quite close to the nominal value, except for one setting for Model 1, where the coverage is still a bit high. As we further increase to samples of 1000 observations, all cases lead to coverage very close to 95%, as we would expect. The simulation standard errors are around 0.002 for all cases, so that for large n most differences in the tables can simply be accounted for by Monte Carlo error. For Model 3, the choice of B (we have also tried $B = 10$ and $B = 30$) did not seem to have any noticeable effect. Overall, the frequentist coverage properties of the models examined are pretty good, with perhaps Model 2 displaying the best performance.

We also conducted the same simulation study using a Student-t sampling model with 2 degrees of freedom and we observed a rather similar behaviour of the coverage proportions.

Appendix C

Proofs for Chapter 5

Proof of Remark 1

Using the fact that the transformation from $(\boldsymbol{\xi_1}, \boldsymbol{\xi_2})$ to θ,

$$\theta = \int_{\mathbb{R}} F_1(y; \boldsymbol{\xi_1}) f_2(y; \boldsymbol{\xi_2}) dy,$$

is a measurable function of the parameters (Enis and Geisser, 1971), we get that the posterior distribution of θ is proper if the posterior of $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ is also proper.

Proof of Corollary 10

(i) follows using the upper bound

$$f_j(x; \mu_j, \sigma_j, \pi_j) \leq \frac{2}{\sigma_j} s_j \left(\frac{x - \mu_j}{\sigma_j} \right), \tag{C.1}$$

together with Remark 1 above and the properness of the posterior in the symmetric case under this prior structure implied by Theorem 1 from Fernández and Steel (1998b).
(ii) follows using this upper bound and Theorem 2 from Fernández and Steel (1998b).

Proof of Theorem 16

Using Remark 1 we have that it suffices to prove existence of the posterior distribution of the model parameters.

For model $(5.2) - (5.3)$, let s_1 be a scale mixture of normals with τ_j the mixing variable associated with x_j and where the τ_j's are independent random variables defined on $\mathbb{R}^+$ with distribution P_{τ_j}. We get an upper bound for the marginal distribution of $\mathbf{x} = (x_1, \ldots, x_{n_1})$ proportional to

$$\int_{S_1\times\cdots\times S_{n_1}} \int_{\mathbb{R}^+_{n_1}} \int_{\Gamma_1} \int_0^\infty \int_{-\infty}^\infty \left(\prod_{j=1}^{n_1} \tau_j^{\frac{1}{2}}\right) \frac{\sigma_1^{-(n_1+1)}}{[a(\gamma_1)+b(\gamma_1)]^{n_1}}$$
$$\times \quad \exp\left[-\frac{1}{2\sigma_1^2 h(\gamma_1)^2}\sum_{j=1}^{n_1}\tau_j(x_j-\mu_1)^2\right] p_{\gamma_1}(\gamma_1)\, d\mu_1 d\sigma_1 d\gamma_1 dP_{(\tau_1,\ldots,\tau_{n_1})} d\mathbf{x},$$

where $h(\gamma_1) = \max\{a(\gamma_1), b(\gamma_1)\}$ and $p_{\gamma_1}(\gamma_1)$ is the factor dependent of γ_1 in (5.3). Consider the change of variable $\vartheta = \sigma_1 h(\gamma_1)$ and rewrite the upper bound as follows

$$\int_{\Gamma_1}\left[\frac{h(\gamma_1)}{a(\gamma_1)+b(\gamma_1)}\right]^{n_1} p_{\gamma_1}(\gamma_1)\, d\gamma_1 \int_{S_1\times\cdots\times S_{n_1}} \int_{\mathbb{R}^+_{n_1}} \int_0^\infty \int_{-\infty}^\infty \left(\prod_{j=1}^{n_1} \tau_j^{\frac{1}{2}}\right) \vartheta^{-(n_1+1)}$$
$$\times \quad \exp\left[-\frac{1}{2\vartheta^2}\sum_{j=1}^{n_1}\tau_j(x_j-\mu_1)^2\right] d\mu_1 d\vartheta dP_{(\tau_1,\ldots,\tau_{n_1})} d\mathbf{x}.$$

The integral with respect to γ_1 is finite for any n_1 and by Theorem 4 from Fernández and Steel (1998b) we have that the integral in $(\mu_1, \vartheta, \tau_1, \ldots, \tau_{n_1}, \mathbf{x})$ is finite if (5.13) is satisfied. Analogously for $\mathbf{y}$.

For model $(5.5) - (5.6)$, using inequality $(C.1)$, we find that for skew-symmetric scale mixtures of normals sampling models the posterior of θ exists whenever the posterior distribution of the parameters in the symmetric case exists. Thus, by Theorem 4 from Fernández and Steel (1998b) this happens whenever (5.13) is satisfied.

Appendix D

Proofs for Chapter 6

Proof of Theorem 17

(i) First, note that by construction we have

$$
\begin{aligned}
f(x_j;\mu,\sigma,\delta_1,\delta_2) &= \int_0^\infty \frac{2\lambda_j^{\frac{1}{2}}}{\sqrt{2\pi}\sigma}\exp\left[-\frac{\lambda_j}{2\sigma^2}(x_j-\mu)^2\right]\\
&\times \left\{\epsilon dP_{\lambda_j|\delta_1} I(x_j<\mu) + (1-\epsilon) dP_{\lambda_j|\delta_2} I(x_j\geq\mu)\right\},
\end{aligned}
$$

with ϵ as in (6.2). Then, we can write the marginal of $\mathbf{x}$ as follows

$$
\begin{aligned}
p(\mathbf{x}) &\propto \int_\Delta\int_\Delta\int_0^\infty\int_{-\infty}^\infty\int_{\mathbb{R}_+^n} \frac{\prod_{j=1}^n \lambda_j^{\frac{1}{2}}}{\sigma^{n+1}}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^n \lambda_j(x_j-\mu)^2\right] p(\delta_1,\delta_2)\\
&\times \prod_{j=1}^n \left\{\epsilon dP_{\lambda_j|\delta_1} I(x_j<\mu) + (1-\epsilon) dP_{\lambda_j|\delta_2} I(x_j\geq\mu)\right\} d\mu d\sigma d\delta_1 d\delta_2.
\end{aligned}
$$

Consider separating the integral with respect to μ into $n+1$ integrals over the domains $(-\infty, x_{(1)})$, $[x(1), x_{(2)})$, ..., $[x_{(n)}, \infty)$, then we have that

$$
\begin{aligned}
I_1 &= \int_\Delta\int_\Delta\int_0^\infty\int_{-\infty}^{x_{(1)}}\int_{\mathbb{R}_+^n} \frac{\prod_{j=1}^n \lambda_j^{\frac{1}{2}}}{\sigma^{n+1}}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^n \lambda_j(x_j-\mu)^2\right] p(\delta_1,\delta_2)\\
&\times (1-\epsilon)^n \prod_{j=1}^n dP_{\lambda_j|\delta_2} d\mu d\sigma d\delta_1 d\delta_2.
\end{aligned}
$$

By noting that $0 \le \epsilon \le 1$, extending the integration domain on μ to the whole real line and integrating out δ_1 we obtain

$$
\begin{aligned}
I_1 &\le \int_{\Delta}\int_{0}^{\infty}\int_{-\infty}^{\infty}\int_{\mathbb{R}_+^n} \frac{\prod_{j=1}^{n}\lambda_j^{\frac{1}{2}}}{\sigma^{n+1}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\lambda_j(x_j-\mu)^2\right] p(\delta_2) \\
&\times \prod_{j=1}^{n} dP_{\lambda_j|\delta_2} d\mu d\sigma d\delta_2 < \infty.
\end{aligned}
$$

The finiteness of this integral is obtained using Theorem 1 from Fernández and Steel (1998b). Now, using similar arguments we have that

$$
\begin{aligned}
I_2 &= \int_{\Delta}\int_{\Delta}\int_{0}^{\infty}\int_{x_{(n)}}^{\infty}\int_{\mathbb{R}_+^n} \frac{\prod_{j=1}^{n}\lambda_j^{\frac{1}{2}}}{\sigma^{n+1}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\lambda_j(x_j-\mu)^2\right] p(\delta_1,\delta_2) \\
&\times \epsilon^n \prod_{j=1}^{n} dP_{\lambda_j|\delta_1} d\mu d\sigma d\delta_1 d\delta_2 \\
&\le \int_{\Delta}\int_{0}^{\infty}\int_{-\infty}^{\infty}\int_{\mathbb{R}_+^n} \frac{\prod_{j=1}^{n}\lambda_j^{\frac{1}{2}}}{\sigma^{n+1}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\lambda_j(x_j-\mu)^2\right] p(\delta_1) \\
&\times \prod_{j=1}^{n} dP_{\lambda_j|\delta_1} d\mu d\sigma d\delta_1 < \infty.
\end{aligned}
$$

Finally, for an intermediate region we have

$$
\begin{aligned}
I_3 &= \int_{\Delta}\int_{\Delta}\int_{0}^{\infty}\int_{x_{(k)}}^{x_{(k+1)}}\int_{\mathbb{R}_+^n} \frac{\prod_{j=1}^{n}\lambda_j^{\frac{1}{2}}}{\sigma^{n+1}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\lambda_j(x_{(j)}-\mu)^2\right] p(\delta_1,\delta_2) \\
&\times \epsilon^k(1-\epsilon)^{n-k} \prod_{j=1}^{k} dP_{\lambda_j|\delta_1} \prod_{j=k+1}^{n} dP_{\lambda_j|\delta_2} d\mu d\sigma d\delta_1 d\delta_2 \\
&\le \int_{\Delta}\int_{\Delta}\int_{0}^{\infty}\int_{-\infty}^{\infty}\int_{\mathbb{R}_+^n} \frac{\prod_{j=1}^{n}\lambda_j^{\frac{1}{2}}}{\sigma^{n+1}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\lambda_j(x_{(j)}-\mu)^2\right] p(\delta_1,\delta_2) \\
&\times \prod_{j=1}^{k} dP_{\lambda_j|\delta_1} \prod_{j=k+1}^{n} dP_{\lambda_j|\delta_2} d\mu d\sigma d\delta_1 d\delta_2 < \infty.
\end{aligned}
$$

The finiteness follows again by Theorem 1 from Fernández and Steel (1998b). Combining the finiteness of I_1, I_2 and I_3 the result follows.

(ii) The result follows using the previous proof, Remark 2 and Theorem 2 from Fernández and Steel (1998b).

Proof of Theorem 18

(i) First of all note that, by using this parameterisation, ϵ in (6.2) does not depend on σ. This fact will be used implicitly in a change of variable below.

Note that

$$
\begin{aligned}
p(\mathbf{x}) \;\propto\;& \int_{\Delta}\int_{\Delta}\int_{0}^{\infty}\int_{-\infty}^{\infty}\int_{\mathbb{R}_+^n} \frac{1}{[a(\gamma)+b(\gamma)]^n}\frac{\prod_{j=1}^{n}\lambda_j^{\frac{1}{2}}}{\sigma^{n+1}} \\
\times\;& \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\frac{\lambda_j}{i_j(\gamma)^2}(x_j-\mu)^2\right] p(\delta_1,\delta_2)p(\gamma) \\
\times\;& \prod_{j=1}^{n}\Big\{\epsilon dP_{\lambda_j|\delta_1} I(x_j<\mu) + (1-\epsilon) dP_{\lambda_j|\delta_2} I(x_j\geq\mu)\Big\}\, d\mu d\sigma d\gamma d\delta_1 d\delta_2 \\
\leq\;& \int_{\Delta}\int_{\Delta}\int_{0}^{\infty}\int_{-\infty}^{\infty}\int_{\mathbb{R}_+^n} \frac{1}{[a(\gamma)+b(\gamma)]^n}\frac{\prod_{j=1}^{n}\lambda_j^{\frac{1}{2}}}{\sigma^{n+1}} \\
\times\;& \exp\left[-\frac{1}{2\sigma^2 h(\gamma)^2}\sum_{j=1}^{n}\lambda_j(x_j-\mu)^2\right] p(\gamma,\delta_1,\delta_2) \\
\times\;& \prod_{j=1}^{n}\Big\{\epsilon dP_{\lambda_j|\delta_1} I(x_j<\mu) + (1-\epsilon) dP_{\lambda_j|\delta_2} I(x_j\geq\mu)\Big\}\, d\mu d\sigma d\gamma d\delta_1 d\delta_2,
\end{aligned}
$$

where $i_j(\gamma) = a(\gamma)I(x_j \geq \mu) + b(\gamma)I(x_j < \mu)$ and $h(\gamma) = \max\{a(\gamma), b(\gamma)\}$. Now, consider the change of variable $\theta = \sigma h(\gamma)$, then we get that this upper bound can be written as follows

$$\int_\Delta \int_\Delta \int_0^\infty \int_{-\infty}^\infty \int_{\mathbb{R}_+^n} \frac{h(\gamma)^n}{[a(\gamma)+b(\gamma)]^n} \frac{\prod_{j=1}^n \lambda_j^{\frac{1}{2}}}{\theta^{n+1}}$$
$$\times \quad \exp\left[-\frac{1}{2\theta^2}\sum_{j=1}^n \lambda_j (x_j-\mu)^2\right] p(\gamma,\delta_1,\delta_2)$$
$$\times \quad \prod_{j=1}^n \Big\{\epsilon dP_{\lambda_j|\delta_1} I(x_j<\mu) + (1-\epsilon) dP_{\lambda_j|\delta_2} I(x_j \geq \mu)\Big\} \, d\mu d\theta d\gamma d\delta_1 d\delta_2 < \infty.$$

The finiteness of this integral follows by using that $0 \leq \epsilon \leq 1$, $\frac{1}{2} \leq \frac{h(\gamma)^n}{[a(\gamma)+b(\gamma)]^n} \leq 1$ and Theorem 17.

(ii) The result follows from the previous proof and Theorem 17.

Appendix E

A Note on the Fisher Information Matrix and Jeffreys Priors of TPSH Distributions

Recall that the TPSH family of distributions is defined as the two-piece density constructed of $f(x;\mu,\sigma,\delta_1)$ truncated at $(-\infty,\mu)$ and $f(x;\mu,\sigma,\delta_2)$ truncated at $[\mu,\infty)$:

$$s(x;\mu,\sigma,\delta_1,\delta_2) = \frac{2\varepsilon}{\sigma} f\left(\frac{x-\mu}{\sigma};\delta_1\right) + \frac{2(1-\varepsilon)}{\sigma} f\left(\frac{x-\mu}{\sigma};\delta_2\right) I(x \geq \mu), \tag{E.1}$$

where

$$\varepsilon = \frac{f(0;\delta_2)}{f(0;\delta_2)+f(0;\delta_1)}.$$

Proposition 1 *The Fisher information matrix of* $(\mu,\sigma,\delta_1,\delta_2)$ *in* $(E.1)$ *is given by*

$$\begin{pmatrix} \frac{C_1(\delta_1,\delta_2)}{\sigma^2} & \frac{C_3(\delta_1,\delta_2)}{\sigma^2} & \frac{C_4(\delta_1,\delta_2)}{\sigma} & \frac{C_5(\delta_1,\delta_2)}{\sigma} \\ \frac{C_3(\delta_1,\delta_2)}{\sigma^2} & \frac{C_2(\delta_1,\delta_2)}{\sigma^2} & \frac{C_6(\delta_1,\delta_2)}{\sigma} & \frac{C_7(\delta_1,\delta_2)}{\sigma} \\ \frac{C_4(\delta_1,\delta_2)}{\sigma} & \frac{C_6(\delta_1,\delta_2)}{\sigma} & K_1(\delta_1,\delta_2) & K_3(\delta_1,\delta_2) \\ \frac{C_5(\delta_1,\delta_2)}{\sigma} & \frac{C_7(\delta_1,\delta_2)}{\sigma} & K_3(\delta_1,\delta_2) & K_2(\delta_1,\delta_2) \end{pmatrix}, \tag{E.2}$$

where the functions $C_1,...,C_6, K1,...,K_3$ *are functions of* (δ_1,δ_2)*, detailed in the proof. The Jeffreys prior and the independence Jeffreys prior are respectively given by*

$$\pi_J(\mu, \sigma, \delta_1, \delta_2) \propto \frac{p(\delta_1, \delta_2)}{\sigma^2}, \tag{E.3}$$

$$\pi_I(\mu, \sigma, \delta_1, \delta_2) \propto \frac{K_1(\delta_1, \delta_2) K_2(\delta_1, \delta_2)}{\sigma}, \tag{E.4}$$

where $p(\delta_1, \delta_2)$ *is a complicated function of* (δ_1, δ_2).

Proof. *The expressions in the Fisher information matrix are simply obtained by calculating the entries*

$$I_{ij} = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i} \log s(x; \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_j} \log s(x; \boldsymbol{\theta})\right)\right],$$

where $\boldsymbol{\theta} = (\mu, \sigma, \delta_1, \delta_2)$. *The functions of* (δ_1, δ_2) *are given by*

$$
\begin{aligned}
C_1(\delta_1,\delta_2) &= \frac{2}{f(0;\delta_1)+f(0;\delta_2)}\left[f(0;\delta_2)\int_{-\infty}^{0}\frac{f'(t;\delta_1)^2}{f(t;\delta_1)}dt\right.\\
&+ \left. f(0;\delta_1)\int_{0}^{\infty}\frac{f'(t;\delta_2)^2}{f(t;\delta_2)}dt\right],\\
C_2(\delta_1,\delta_2) &= \frac{2}{f(0;\delta_1)+f(0;\delta_2)}\left\{f(0;\delta_2)\int_{-\infty}^{0}\left[1+t\frac{f'(t;\delta_1)}{f(t;\delta_1)}\right]^2 f(t;\delta_1)dt\right.\\
&+ \left. f(0;\delta_1)\int_{0}^{\infty}\left[1+t\frac{f'(t;\delta_2)}{f(t;\delta_2)}\right]^2 f(t;\delta_2)dt\right\},\\
C_3(\delta_1,\delta_2) &= \frac{2}{f(0;\delta_1)+f(0;\delta_2)}\left[f(0;\delta_2)\int_{-\infty}^{0}t\frac{f'(t;\delta_1)^2}{f(t;\delta_1)}dt\right.\\
&+ \left. f(0;\delta_1)\int_{0}^{\infty}t\frac{f'(t;\delta_2)^2}{f(t;\delta_2)}dt\right],\\
C_4(\delta_1,\delta_2) &= \frac{2f(0;\delta_2)}{f(0;\delta_1)+f(0;\delta_2)}\left[\dot{f}(0;\delta_1)-\int_{-\infty}^{0}f'(t;\delta_1)\frac{\dot{f}(t;\delta_1)}{f(t;\delta_1)}dt\right],\\
C_5(\delta_1,\delta_2) &= -\frac{2f(0;\delta_1)}{f(0;\delta_1)+f(0;\delta_2)}\left[\dot{f}(0;\delta_2)+\int_{0}^{\infty}f'(t;\delta_2)\frac{\dot{f}(t;\delta_2)}{f(t;\delta_2)}dt\right],\\
C_6(\delta_1,\delta_2) &= -\frac{2f(0;\delta_2)}{f(0;\delta_1)+f(0;\delta_2)}\int_{-\infty}^{0}\dot{f}(t;\delta_1)\left[1+t\frac{f'(t;\delta_1)}{f(t;\delta_1)}\right]dt,\\
C_7(\delta_1,\delta_2) &= -\frac{2f(0;\delta_1)}{f(0;\delta_1)+f(0;\delta_2)}\int_{0}^{\infty}\dot{f}(t;\delta_2)\left[1+t\frac{f'(t;\delta_2)}{f(t;\delta_2)}\right]dt,\\
K_1(\delta_1,\delta_2) &= \frac{1}{f(0;\delta_1)+f(0;\delta_2)}\left\{2\int_{-\infty}^{0}\left[-\frac{\dot{f}(0;\delta_1)}{f(0;\delta_1)+f(0;\delta_2)}+\frac{\dot{f}(t;\delta_1)}{f(t;\delta_1)}\right]^2\right.\\
&\times \left. f(t;\delta_1)dt+\left[-\frac{\dot{f}(0;\delta_1)}{f(0;\delta_1)+f(0;\delta_2)}+\frac{\dot{f}(0;\delta_1)}{f(0;\delta_1)}\right]^2\right\},\\
K_2(\delta_1,\delta_2) &= \frac{1}{f(0;\delta_1)+f(0;\delta_2)}\left\{2\int_{0}^{\infty}\left[-\frac{\dot{f}(0;\delta_2)}{f(0;\delta_1)+f(0;\delta_2)}+\frac{\dot{f}(t;\delta_2)}{f(t;\delta_2)}\right]^2\right.\\
&\times \left. f(t;\delta_2)dt+\left[-\frac{\dot{f}(0;\delta_2)}{f(0;\delta_1)+f(0;\delta_2)}+\frac{\dot{f}(0;\delta_2)}{f(0;\delta_2)}\right]^2\right\},\\
K_3(\delta_1,\delta_2) &= \frac{1}{f(0;\delta_1)+f(0;\delta_2)}\left\{f(0;\delta_1)\dot{f}(0;\delta_2)\int_{-\infty}^{0}\dot{f}(t;\delta_1)dt\right.\\
&+ \left. f(0;\delta_2)\dot{f}(0;\delta_1)\int_{0}^{\infty}\dot{f}(t;\delta_2)dt-\frac{\dot{f}(0;\delta_1)+\dot{f}(0;\delta_2)}{2}\right\},
\end{aligned}
$$

where $f'(x;\delta) = \dfrac{\partial}{\partial t} f(t;\delta)\Big|_{t=x}$ *and* $\dot{f}(x;\delta) = \dfrac{\partial}{\partial d} f(t;d)\Big|_{d=\delta}$

Since the entries of the Fisher information matrix involve complicated functions of (δ_1, δ_2), it is difficult to come up with interpretable conclusions about the models that produce parameter orthogonality. As a consequence, there seems to be no apparent reparameterisation to induce such property.

Using Theorem 17 it is possible to provide sufficient conditions for the existence of the posterior distribution under the use of the independence Jeffreys prior. Although these conditions are difficult to check in practice, the result is presented below for completeness.

Corollary 11 *Let* $(x_1, \dots, x_n)$ *be a sample from (E.1) and assume that* f *is a scale mixture of normals. It follows that the posterior distribution of* $(\mu, \sigma, \delta_1, \delta_2)$ *using the prior* $(E.4)$ *is proper if the product* $K_1(\delta_1, \delta_2)K_2(\delta_1, \delta_2)$ *is integrable on* Δ^2*,* $n \geq 2$*, and all the observations are different.*

Appendix F

On a Subclass of DTP Transformations

In this section we present a subclass of DTP transformations that produces distributions with different shapes but equal mass cumulated on each side of the mode. We employ this transformation to produce a generalised skew-t distribution.

F.0.1 Proposed Transformation

In Chapter 6 we defined the DTP class of transformations through the density

$$s(x;\mu,\sigma_1,\sigma_2,\delta_1,\delta_2) = \frac{2\varepsilon}{\sigma_1} f\left(\frac{x-\mu}{\sigma_1};\delta_1\right) I(x<\mu) + \frac{2(1-\varepsilon)}{\sigma_2} f\left(\frac{x-\mu}{\sigma_2};\delta_2\right) I(x\geq\mu),$$

where $\mu \in \mathbb{R}$; $\sigma_1, \sigma_2 \in \mathbb{R}_+$; $\delta_1, \delta_2 \in \Delta$; $f \in \mathcal{F}$ and

$$\varepsilon = \frac{\sigma_1 f(0;\delta_2)}{\sigma_1 f(0;\delta_2) + \sigma_2 f(0;\delta_1)}.$$

Now, consider the subclass of transformations obtained by fixing $\sigma_1 = \sigma$ and $\sigma_2 = \dfrac{f(0;\delta_2)}{f(0;\delta_1)}\sigma$. The corresponding transformed density is then given by

$$\begin{aligned} s(x;\mu,\sigma,\delta_1,\delta_2) &= \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma};\delta_1\right) I(x<\mu) \\ &+ \frac{1}{\sigma}\frac{f(0;\delta_1)}{f(0;\delta_2)} f\left(\frac{x-\mu}{\sigma}\frac{f(0;\delta_1)}{f(0;\delta_2)};\delta_2\right) I(x\geq\mu), \end{aligned} \tag{F.1}$$

This density is unimodal with mode and location parameter μ, scale parameter σ, and shape parameters (δ_1, δ_2). The original model $f(\cdot;\delta)$ is a particular case of (F.1) for $\delta_1 = \delta_2 = \delta$. This new transformation produces a flexible distribution since the parameters (δ_1, δ_2) control the shape of s on each side of the mode. Although the AG measure of skewness of this model is 0, density (F.1) is not symmetric in the strict sense since $s(\mu + x;\mu,\sigma,\delta_1,\delta_2) \neq s(\mu - x;\mu,\sigma,\delta_1,\delta_2)$ for $\delta_1 \neq \delta_2$. In the next section we explore the combination of this transformation with a known skewing mechanism in order to produce a richer family of distributions containing members with different mass cumulated on each side of the mode.

F.0.2 A Skew-t Distribution

Let $f(\cdot;\delta)$ in (F.1) be the Student's-t distribution with δ degrees of freedom and consider the skewing mechanism proposed in Jones and Pewsey (2009), $h(x;\gamma) = \sinh[\operatorname{arcsinh}(x)-\gamma]$, $\gamma \in \mathbb{R}$. Define the following generalised skew-t distribution based on transforming (F.1) as in Rosco et al. (2011)

$$\begin{aligned} s_2(x;\mu,\sigma,\gamma,\delta_1,\delta_2) &= \frac{1}{\sigma} s\left(h\left(\frac{x-\mu}{\sigma};\gamma\right);0,1,\delta_1,\delta_2\right) \\ &\times \frac{\cosh\left[\operatorname{arcsinh}\left(\frac{x-\mu}{\sigma}\right)-\gamma\right]}{\sqrt{1+\left(\frac{x-\mu}{\sigma}\right)^2}} \end{aligned} \tag{F.2}$$

This distribution can be seen as a generalisation of the skew-t proposed in Rosco et al. (2011), which is contained as a particular case ($\delta_1 = \delta_2$). In addition, this model presents similar features to the skew-t distributions proposed in Jones and Faddy (2003) and Aas and Haff (2006) since it can capture different tail behaviour in each direction for $\delta_1 \neq \delta_2$. Although this distribution is bimodal when the degrees of freedom (δ_1, δ_2) are small ($\delta_1, \delta_2 < 0.1$), these cases may be of little practical interest (Rosco et al., 2011). Figure $F.1$ shows the shapes obtained with different values of the parameters. We leave the study and application of this distribution as a line for future research.

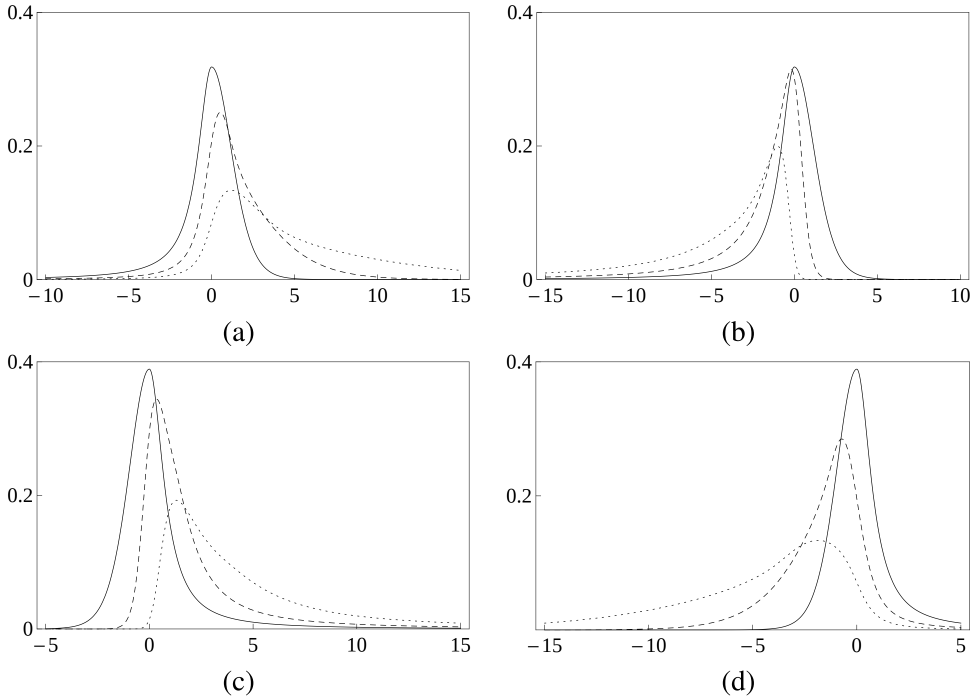


Figure F.1: Shape of density (F.2) for $(\mu, \sigma) = (0, 1)$ and: (a) $(\delta_1, \delta_2) = (1, 10)$, $\gamma = 0, 1, 2$; (b) $(\delta_1, \delta_2) = (1, 10)$, $\gamma = 0, -1, -2$; (c) $(\delta_1, \delta_2) = (10, 1)$, $\gamma = 0, 1, 2$; (d) $(\delta_1, \delta_2) = (10, 1)$, $\gamma = 0, -1, -2$.

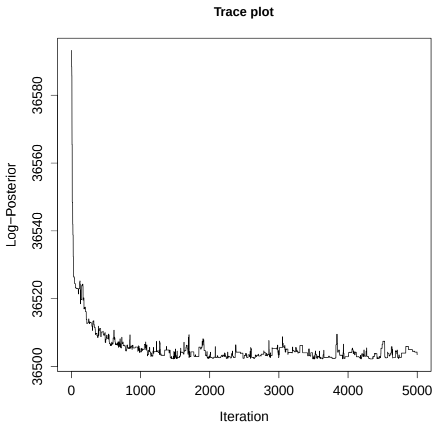
Appendix G

Trace Plots

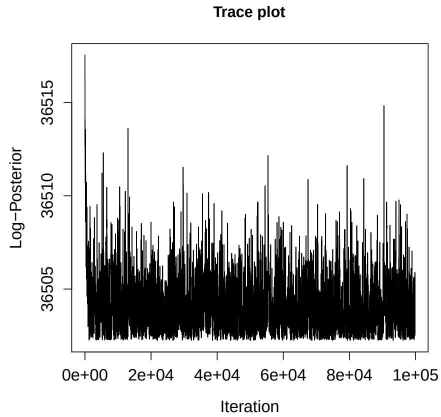
In this appendix we present trace plots of the log–posterior distribution corresponding to the MCMC sampling methods implemented in the examples of Chapters 3–6. For each of these examples we illustrate the convergence of the Markov chain with the trace plot of the first 5,000 iterations. In order to illustrate the long–term behaviour of the Markov chain, we also present the first 100,000 iterations (after removing the first couple of hundred iterations in order to show only the stable part of the trace plot). In addition to this, we present trace plots of the marginal chains of the parameters for some representative examples. Overall, we can observe that the burn–in periods employed throughout the thesis are very conservative given that the stability of the chains is typically reached after a couple of hundred iterations.

G.1 Trace plots for Chapter 3

G.1.1 Section 3.3



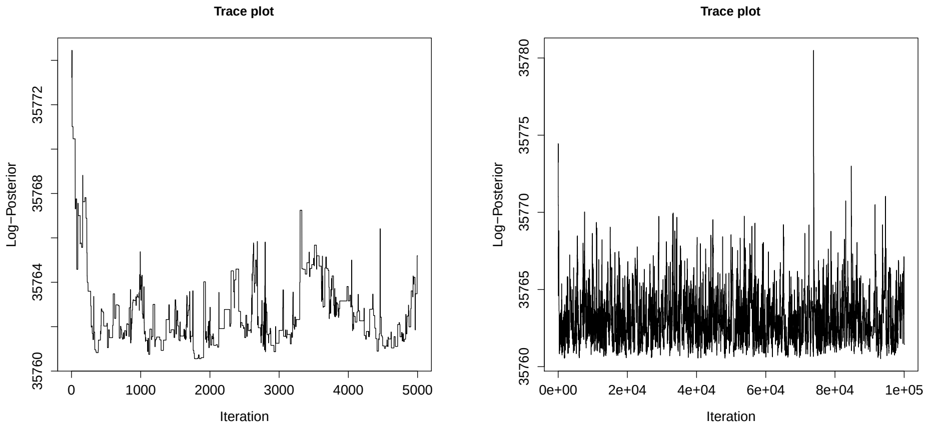
(a)



(b)

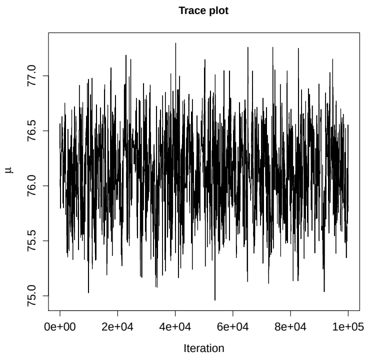
Figure G.1: Untruncated model: (a) First 5,000 iterations of the Log–posterior; (b) First 100,000 Log–posterior.

G.1.2 Section 3.4

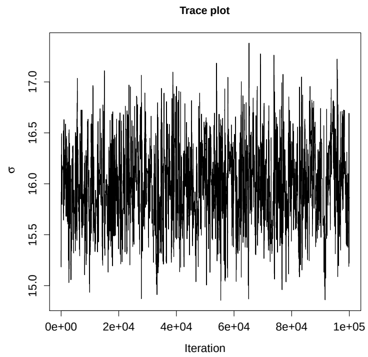


(a) (b)

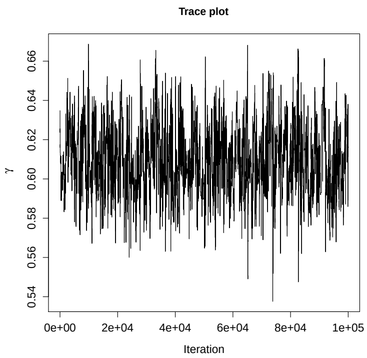
Figure G.2: Doubly–truncated model: (a) First 5,000 iterations of the Log–posterior; (b) First 100,000 Log–posterior.



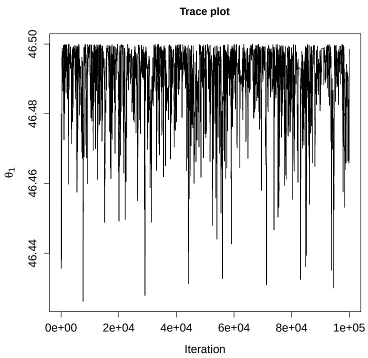
(a)



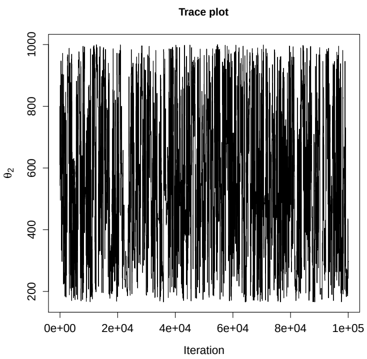
(b)



(c)



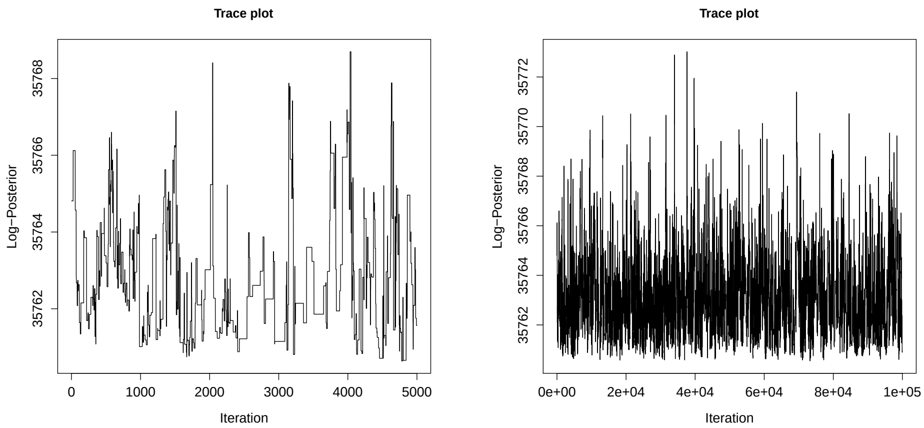
(d)



(e)

Figure G.3: Doubly–truncated model: (a) μ; (b) σ; (c) γ; (d) θ_1; (e) θ_2.

G.1.3 Section 3.5



(a) (b)

Figure G.4: Left–truncated model: (a) First 5,000 iterations of the Log–posterior; (b) First 100,000 Log–posterior.

G.2 Trace plots for Chapter 4

G.2.1 Section 4.4

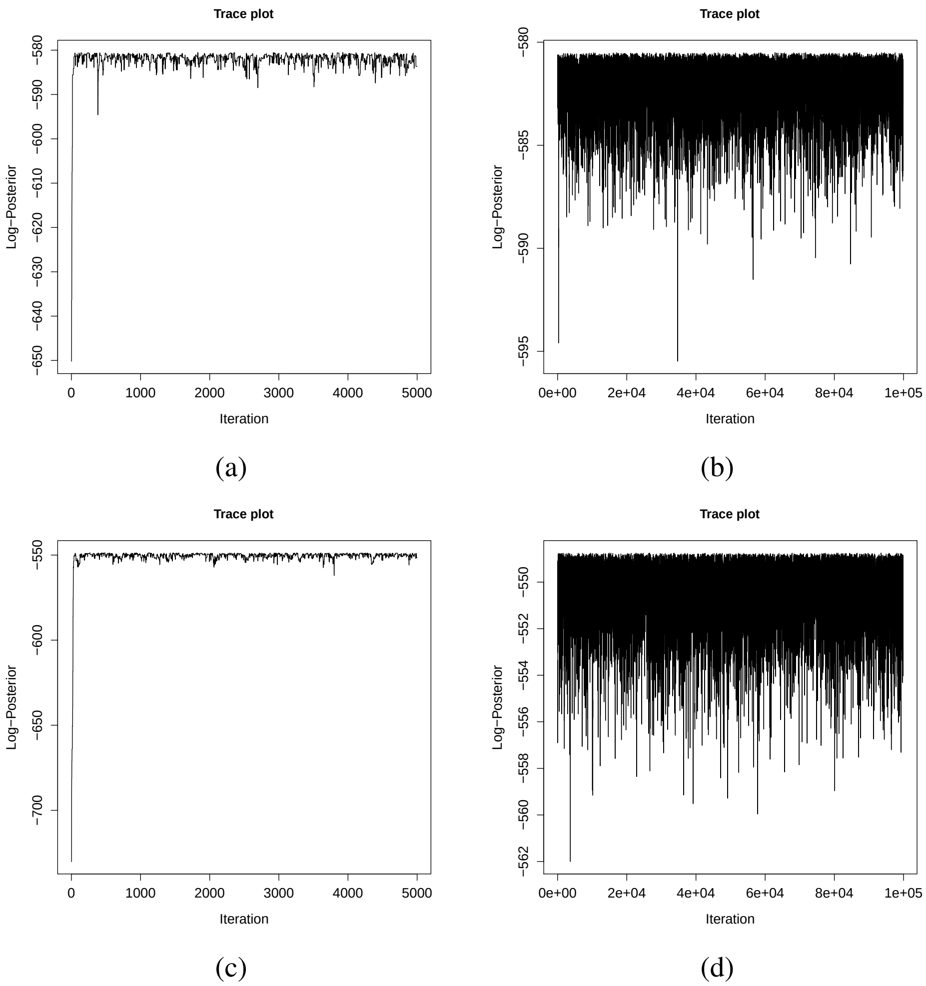


Figure G.5: Model 1: (a) First 5,000 iterations of the Log–posterior (Females); (b) First 100,000 Log–posterior (Females); (c) First 5,000 iterations of the Log–posterior (Males); (d) First 100,000 Log–posterior (Males).

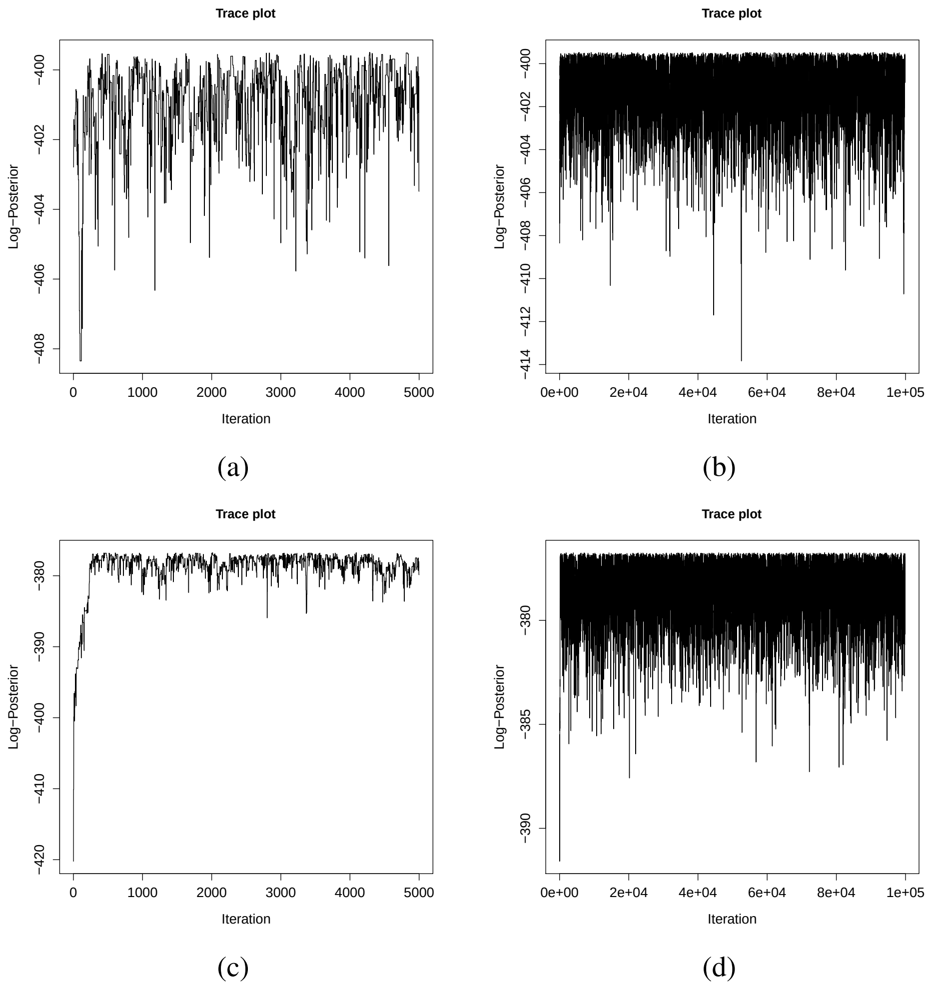
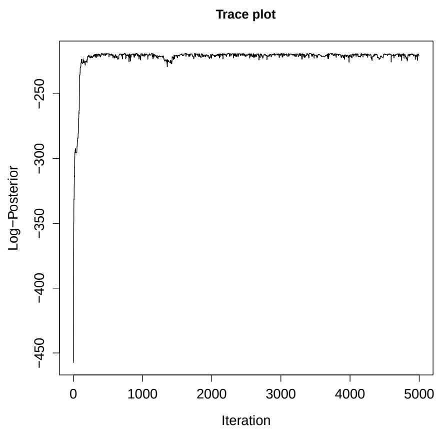
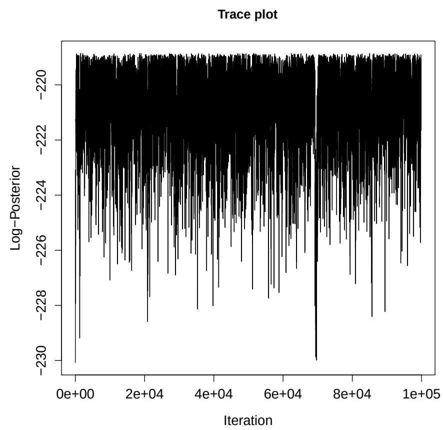


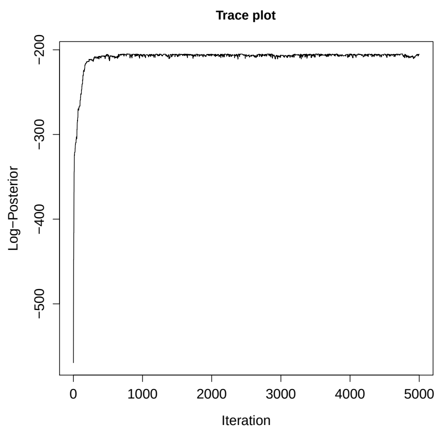
Figure G.6: Model 2: (a) First 5,000 iterations of the Log–posterior (Females); (b) First 100,000 Log–posterior (Females); (c) First 5,000 iterations of the Log–posterior (Males); (d) First 100,000 Log–posterior (Males).



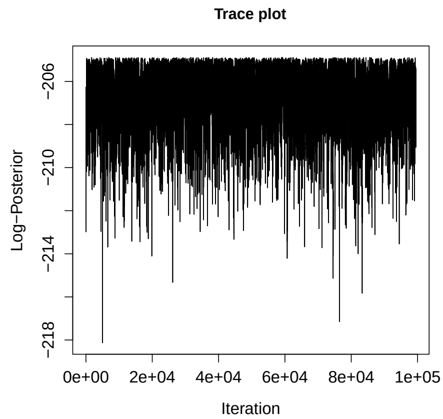
(a)



(b)



(c)



(d)

Figure G.7: Model 3: (a) First 5,000 iterations of the Log–posterior (Females); (b) First 100,000 Log–posterior (Females); (c) First 5,000 iterations of the Log–posterior (Males); (d) First 100,000 Log–posterior (Males).

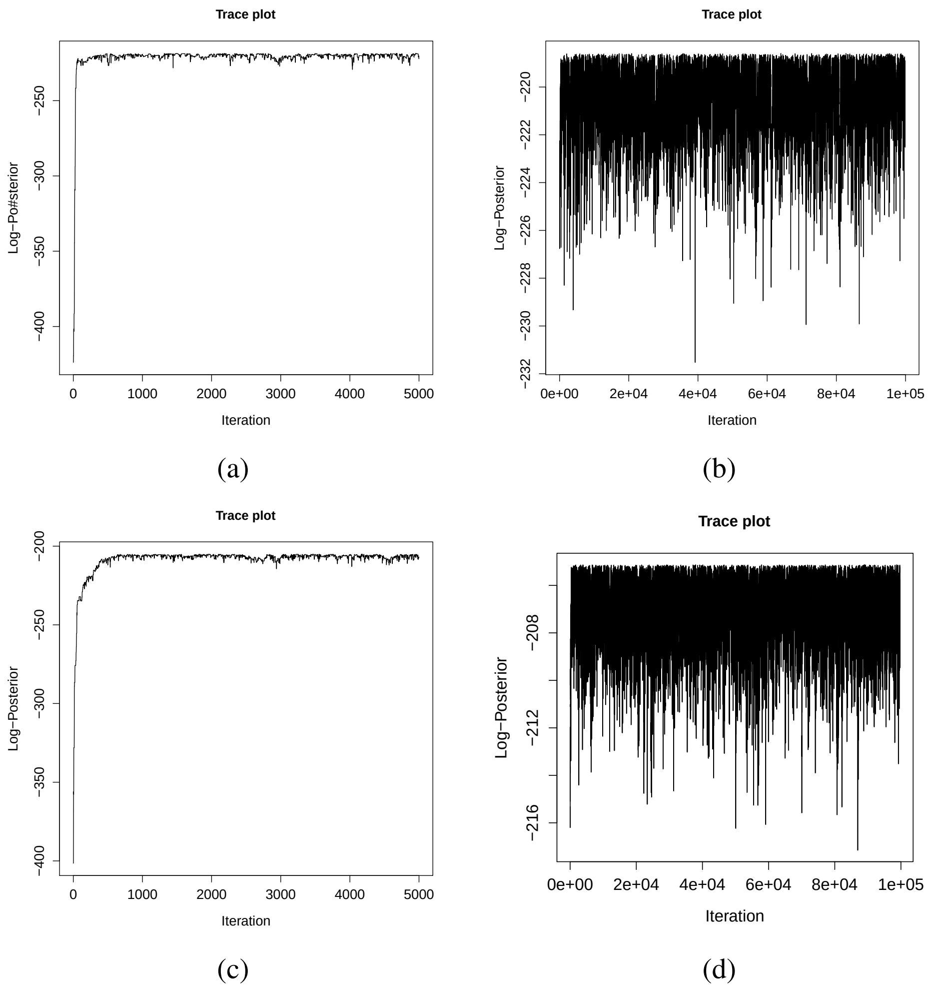


Figure G.8: Model 4: (a) First 5,000 iterations of the Log–posterior (Females); (b) First 100,000 Log–posterior (Females); (c) First 5,000 iterations of the Log–posterior (Males); (d) First 100,000 Log–posterior (Males).

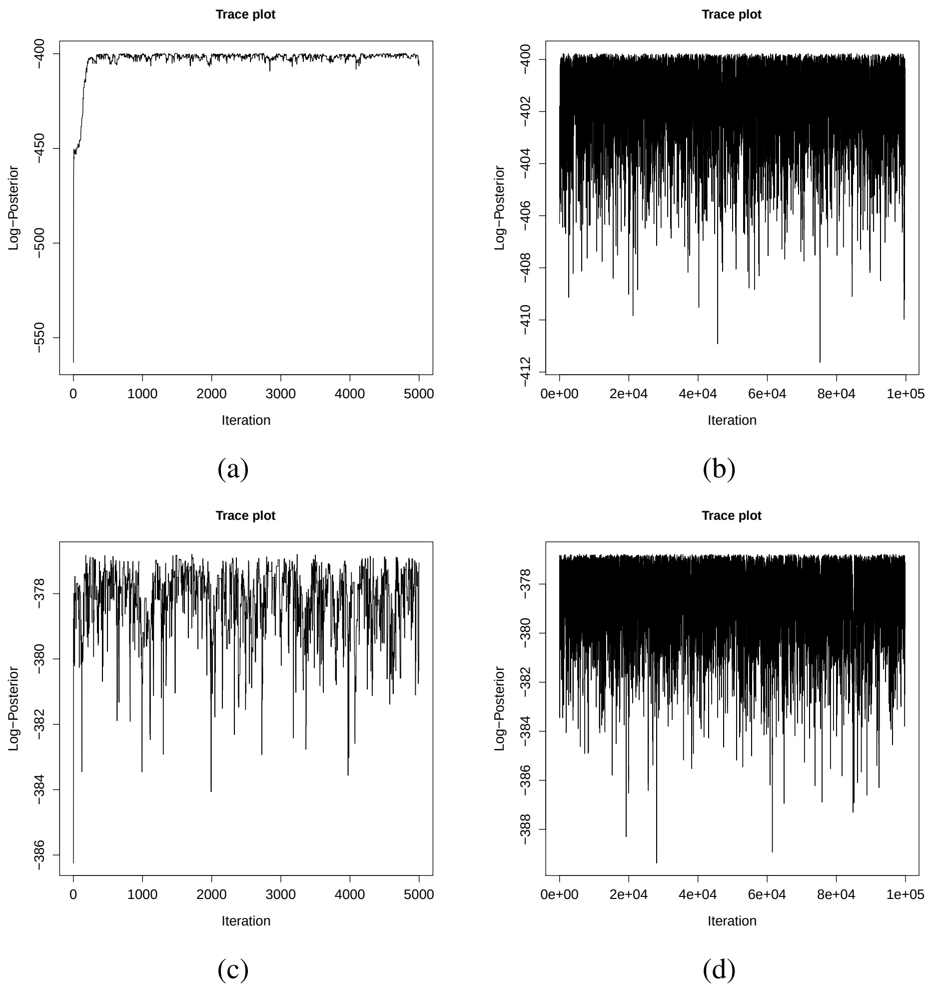


Figure G.9: Model 5: (a) First 5,000 iterations of the Log–posterior (Females); (b) First 100,000 Log–posterior (Females); (c) First 5,000 iterations of the Log–posterior (Males); (d) First 100,000 Log–posterior (Males).

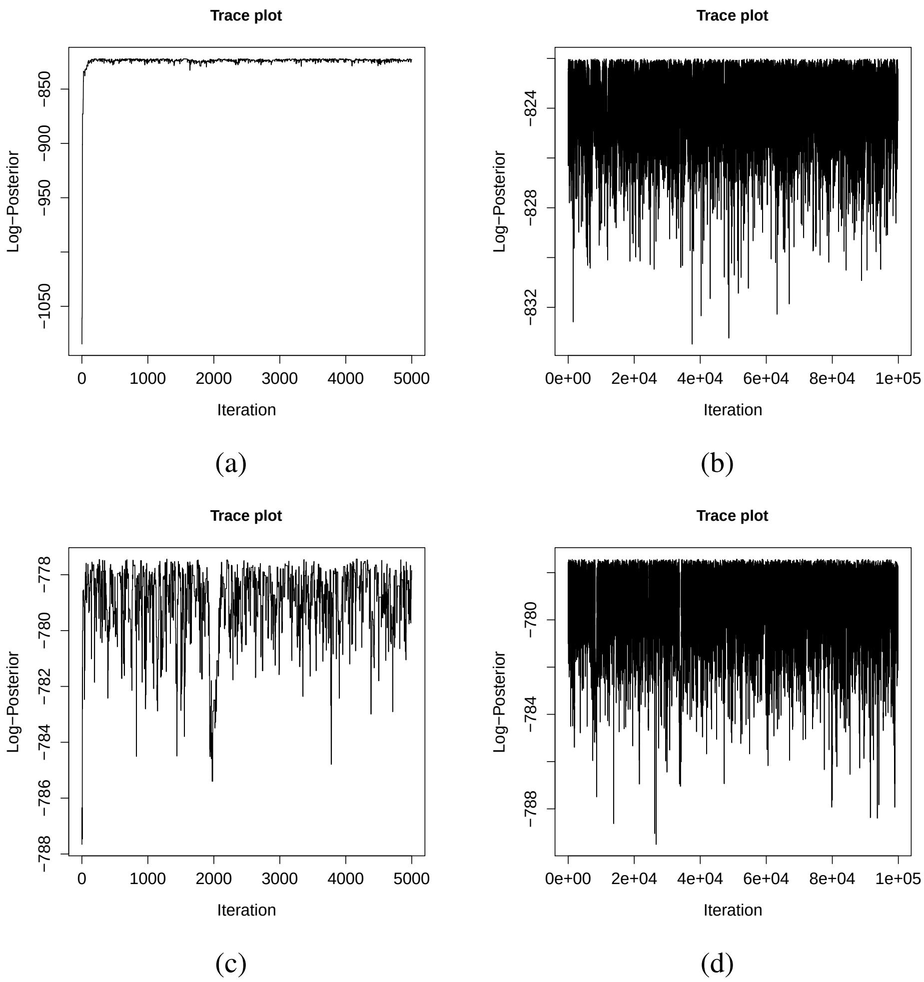


Figure G.10: Model 6: (a) First 5,000 iterations of the Log–posterior (Females); (b) First 100,000 Log–posterior (Females); (c) First 5,000 iterations of the Log–posterior (Males); (d) First 100,000 Log–posterior (Males).

G.3 Trace plots for Chapter 5

G.3.1 Section 5.5.1

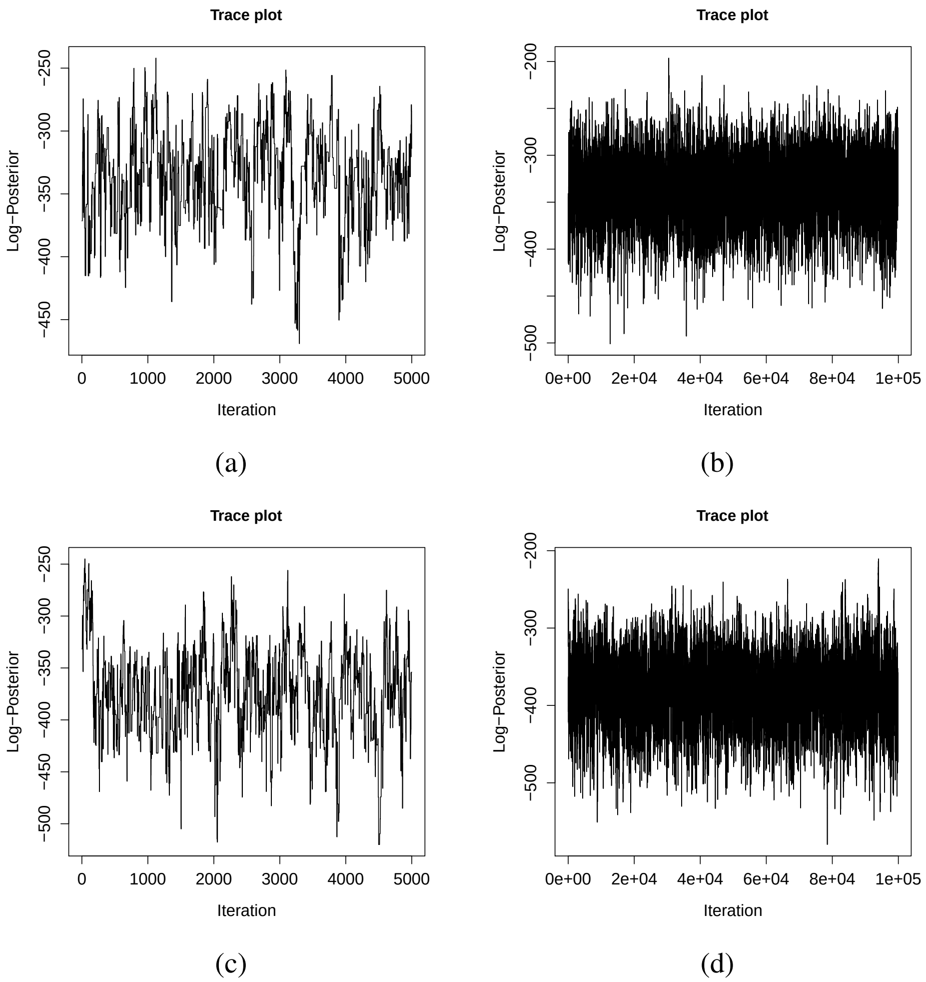


Figure G.11: Simulated data, Normal model: (a) First 5,000 iterations of the Log–posterior (X); (b) First 100,000 Log–posterior (X); (c) First 5,000 iterations of the Log–posterior (Y); (d) First 100,000 Log–posterior (Y).

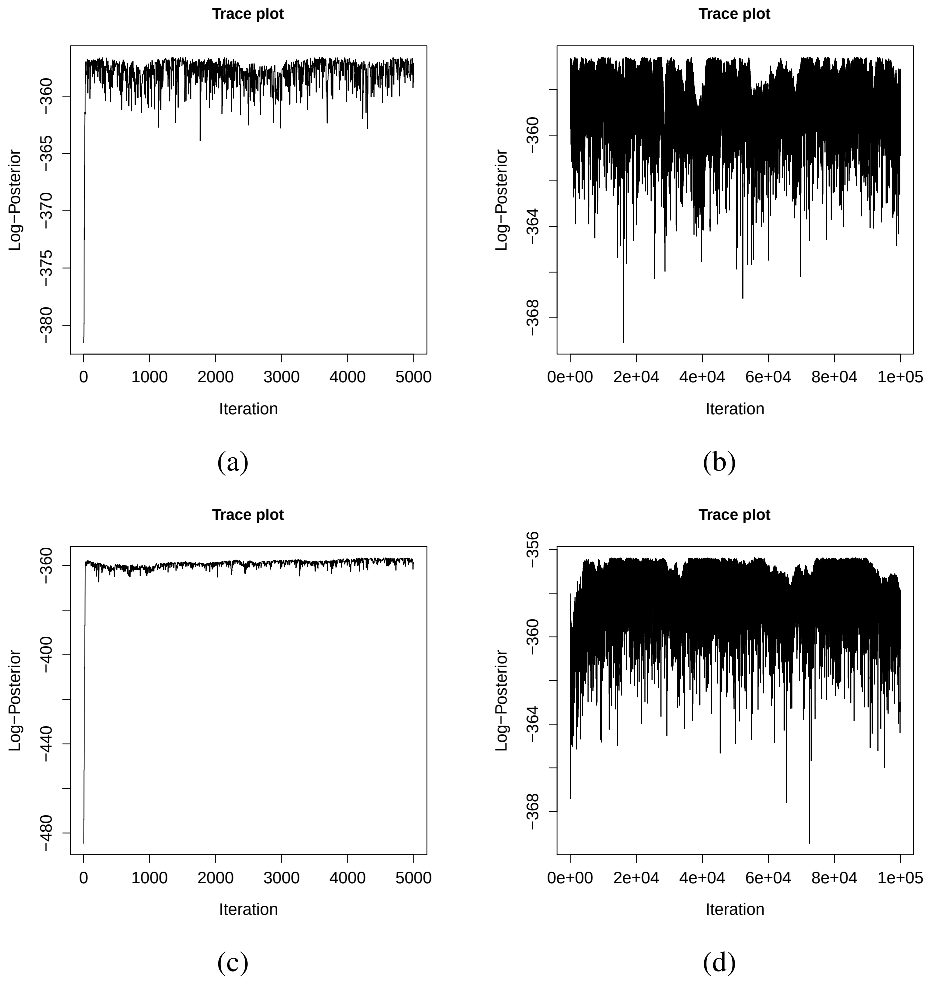


Figure G.12: Simulated data, skew–normal model: (a) First 5,000 iterations of the Log–posterior (X); (b) First 100,000 Log–posterior (X); (c) First 5,000 iterations of the Log–posterior (Y); (d) First 100,000 Log–posterior (Y).

G.3.2 Section 5.5.1

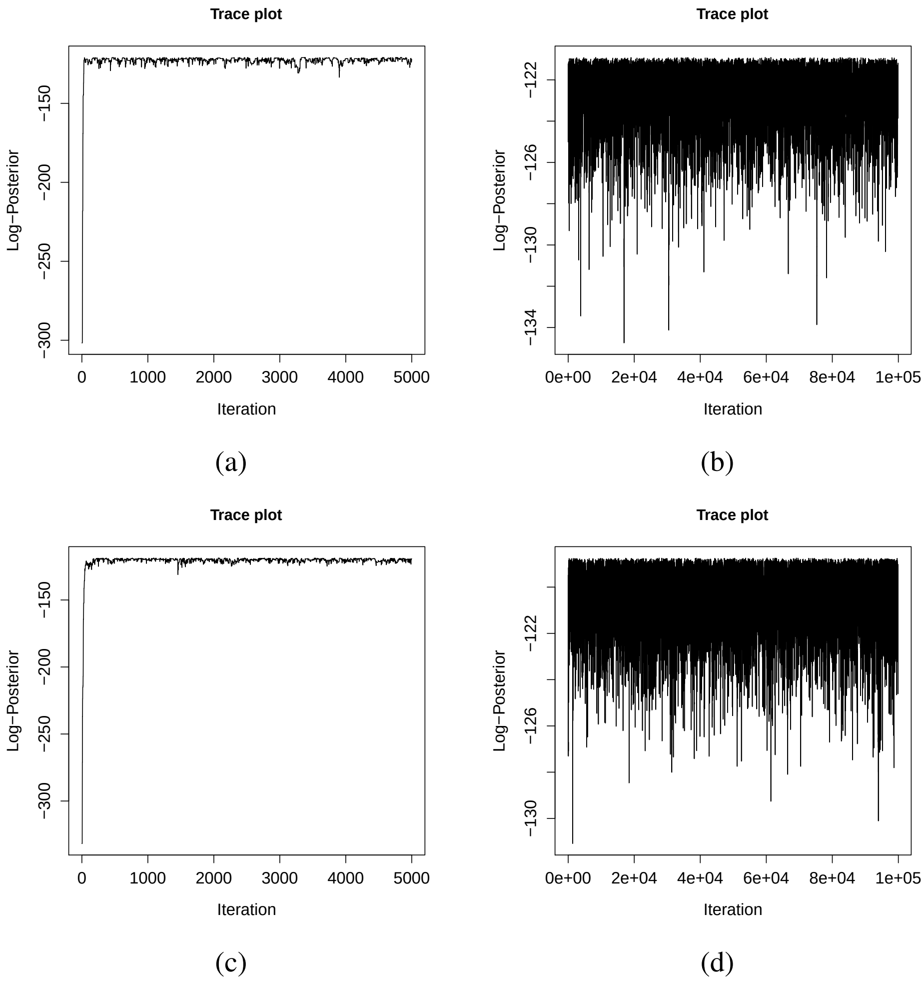


Figure G.13: Simulated data, two–piece normal model: (a) First 5,000 iterations of the Log–posterior (X); (b) First 100,000 Log–posterior (X); (c) First 5,000 iterations of the Log–posterior (Y); (d) First 100,000 Log–posterior (Y).

G.3.3 Section 5.5.2

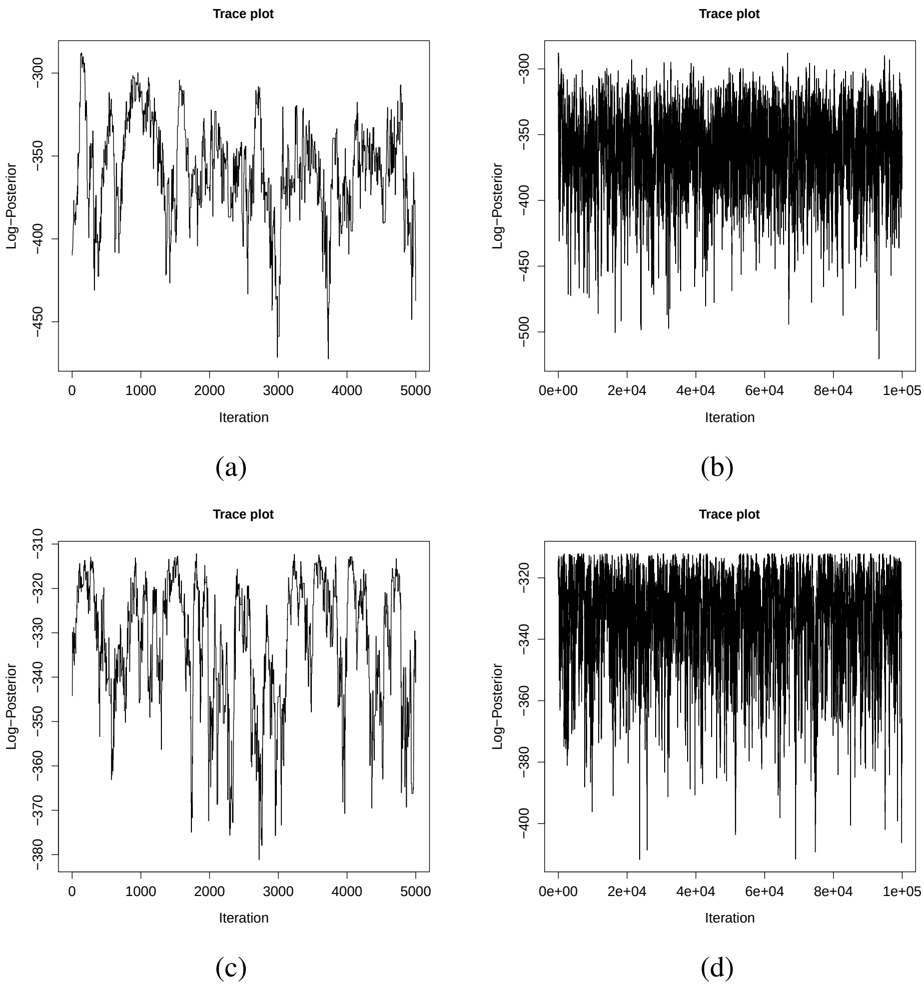


Figure G.14: Body measurements data, Normal model: (a) First 5,000 iterations of the Log–posterior (Females); (b) First 100,000 Log–posterior (Females); (c) First 5,000 iterations of the Log–posterior (Males); (d) First 100,000 Log–posterior (Males).

G.3.4 Section 5.5.3

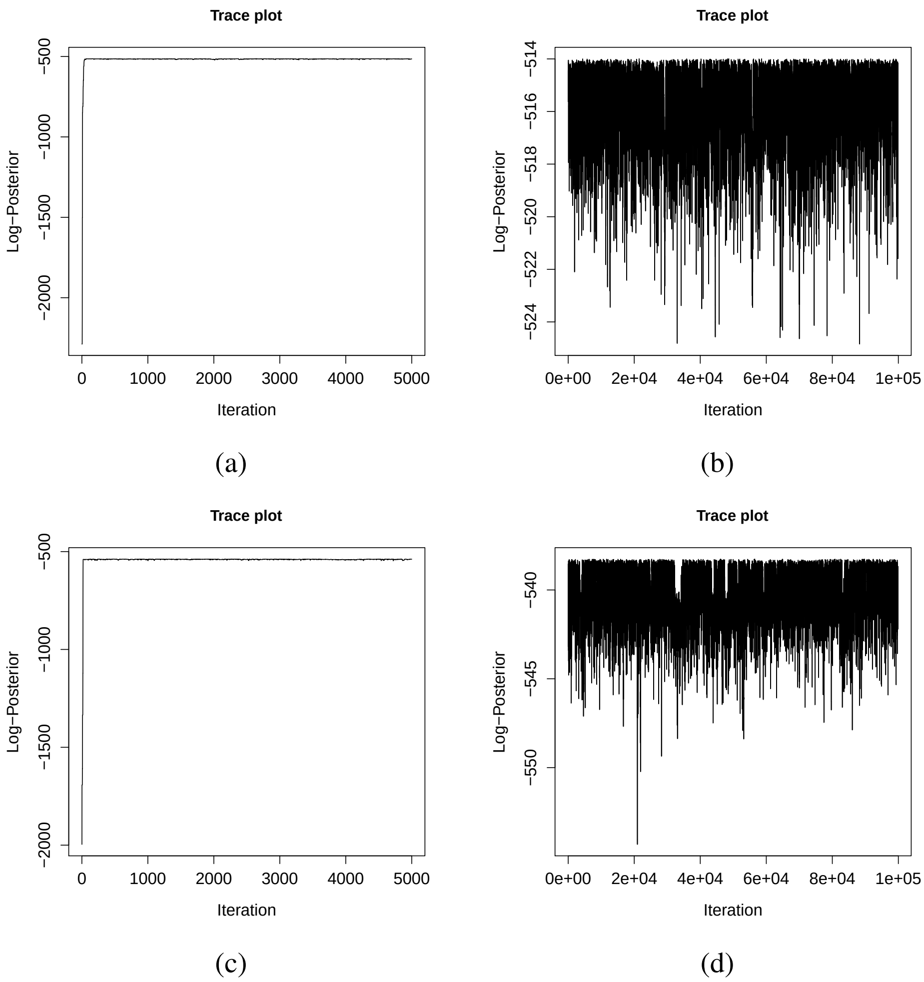


Figure G.15: Body measurements data, skew–normal model: (a) First 5,000 iterations of the Log–posterior (Females); (b) First 100,000 Log–posterior (Females); (c) First 5,000 iterations of the Log–posterior (Males); (d) First 100,000 Log–posterior (Males).

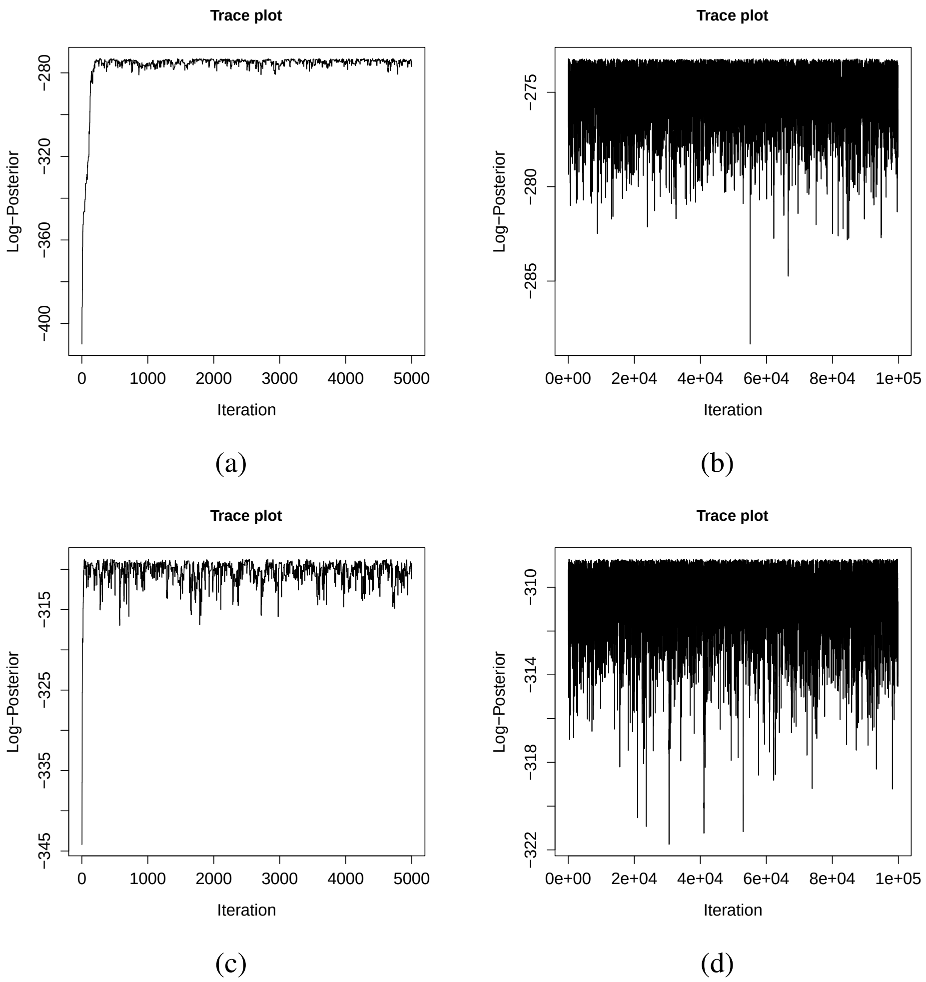


Figure G.16: Body measurements data, two–piece normal model: (a) First 5,000 iterations of the Log–posterior (Females); (b) First 100,000 Log–posterior (Females); (c) First 5,000 iterations of the Log–posterior (Males); (d) First 100,000 Log–posterior (Males).

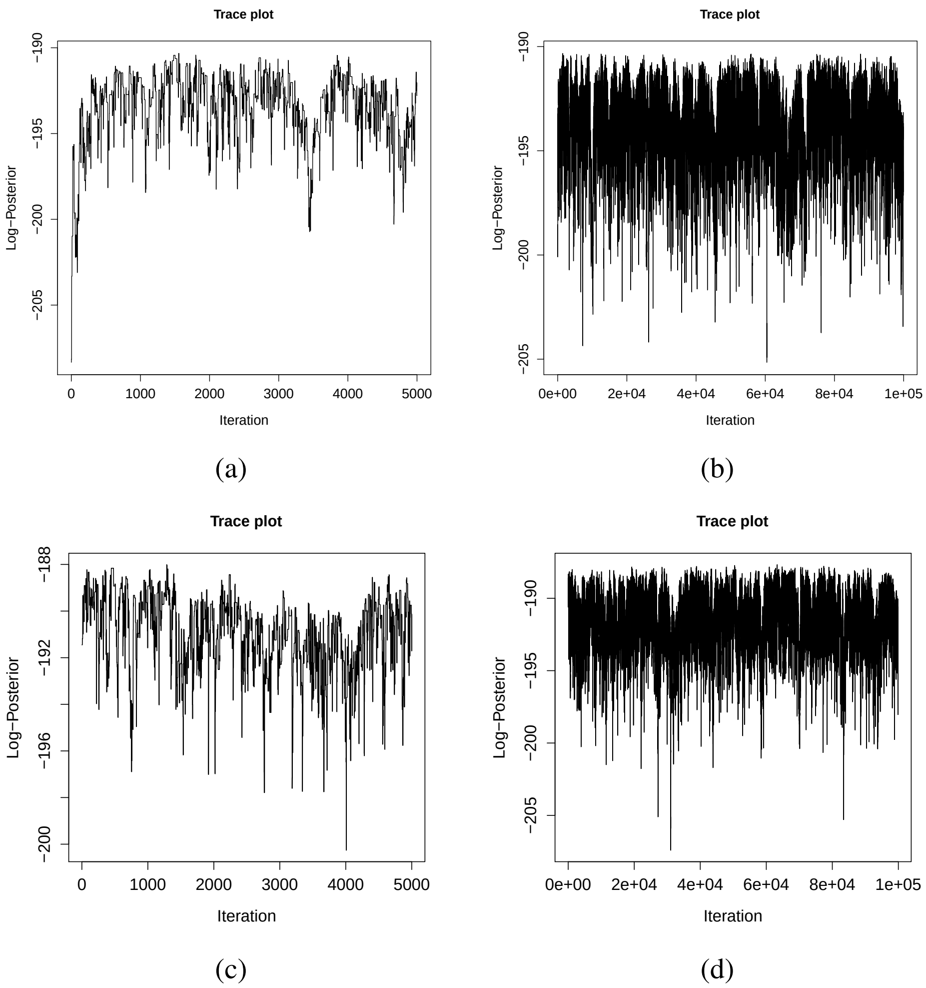


Figure G.17: Melanoma data, dependent model: (a) First 5,000 iterations of the Log–posterior (skew–normal); (b) First 100,000 Log–posterior (skew–normal); (c) First 5,000 iterations of the Log–posterior (two–piece normal); (d) First 100,000 Log–posterior (two–piece normal).

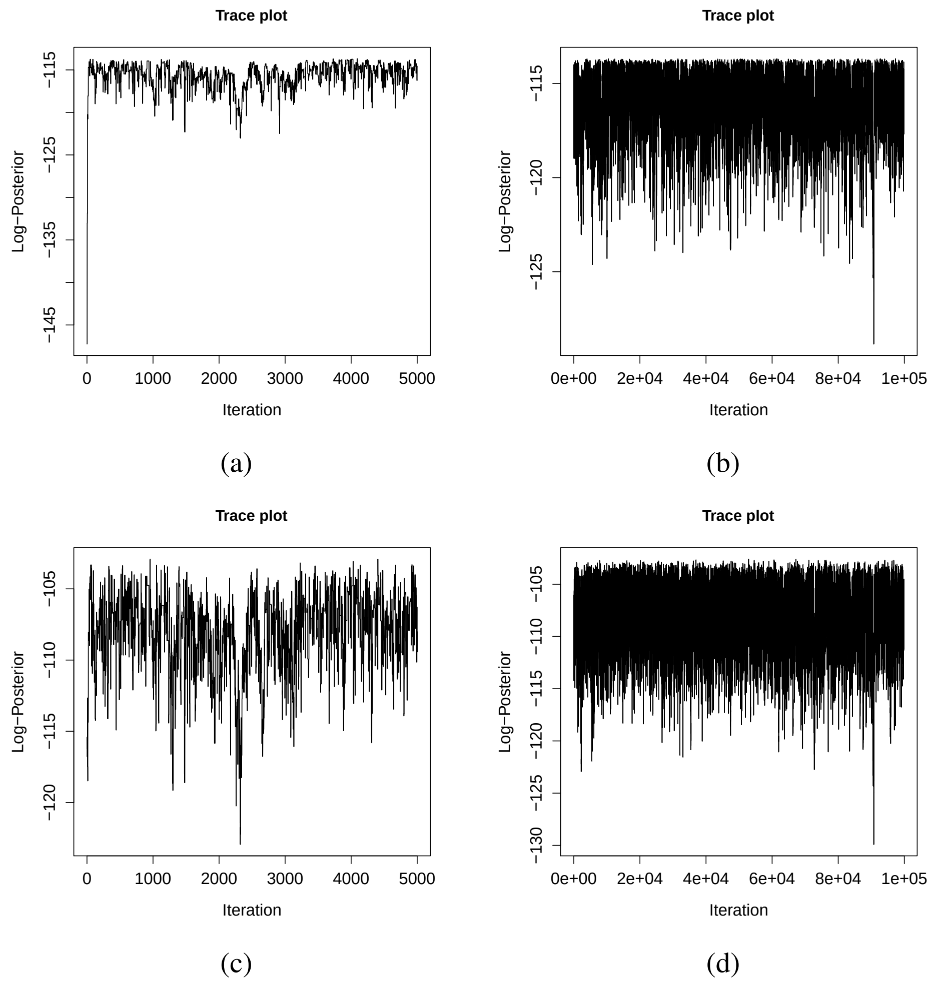


Figure G.18: Melanoma data, skew–normal independent model: (a) First 5,000 iterations of the Log–posterior (X test); (b) First 100,000 Log–posterior (X test); (c) First 5,000 iterations of the Log–posterior (Y test); (d) First 100,000 Log–posterior (Y test).

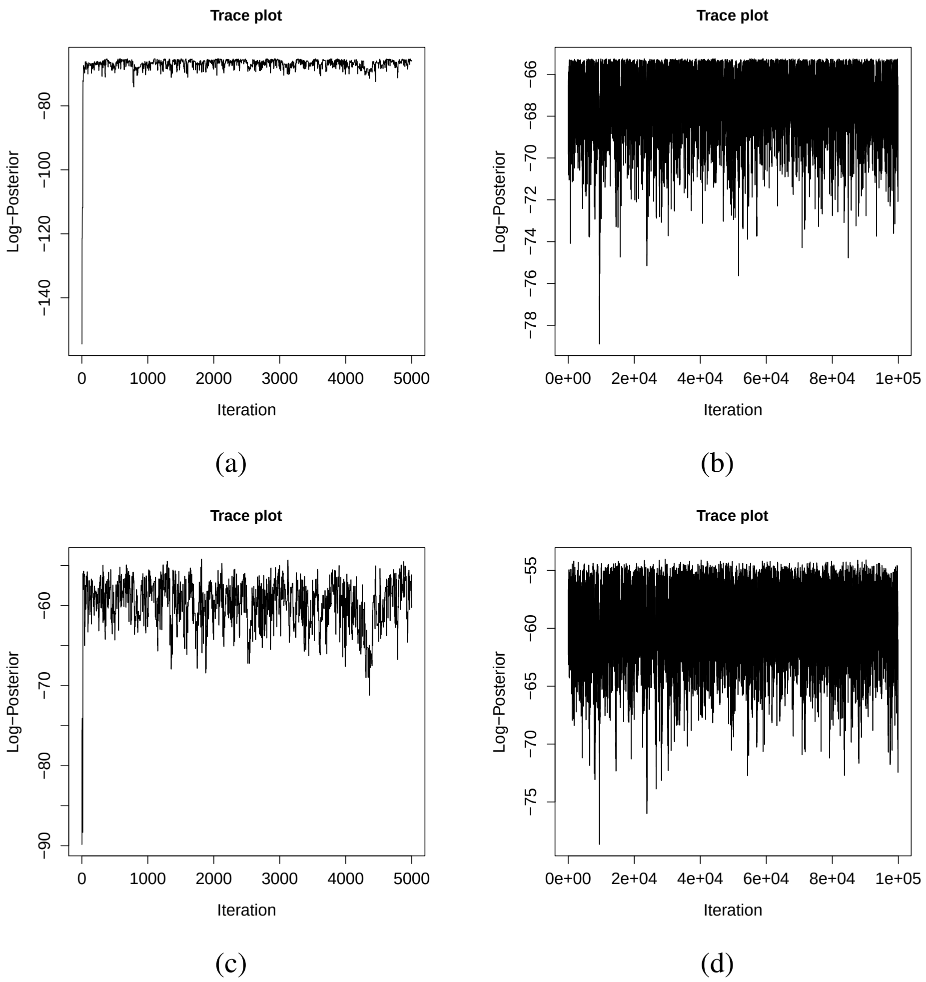


Figure G.19: Melanoma data, two–piece normal independent model: (a) First 5,000 iterations of the Log–posterior (X test); (b) First 100,000 Log–posterior (X test); (c) First 5,000 iterations of the Log–posterior (Y test); (d) First 100,000 Log–posterior (Y test).

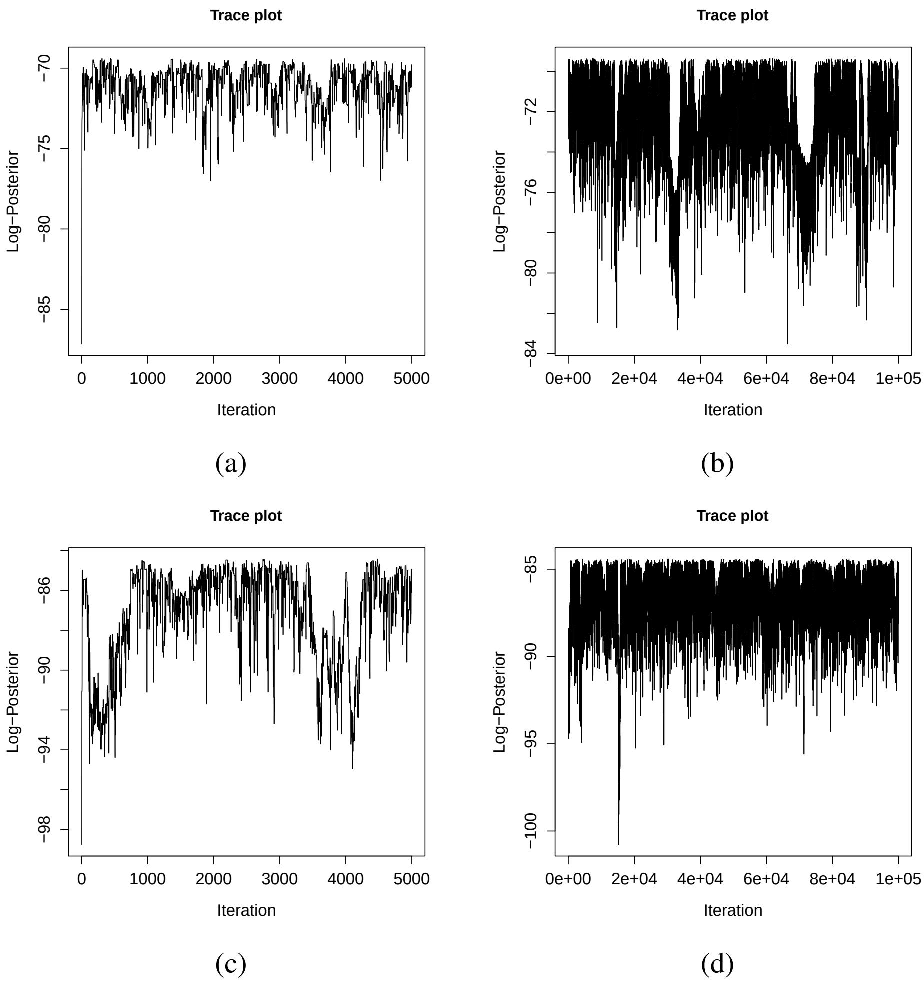
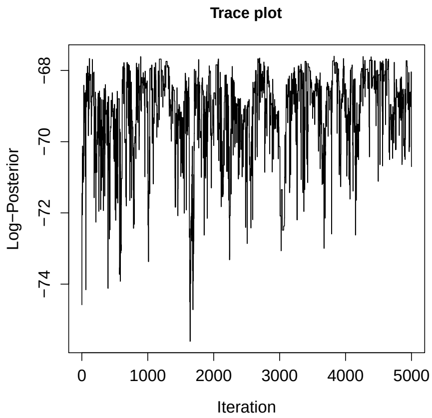
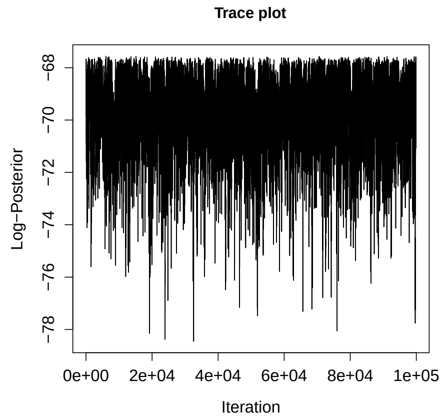


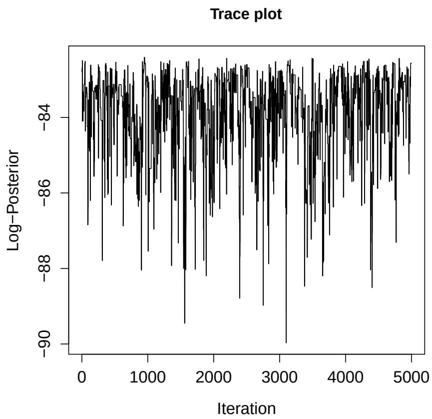
Figure G.20: Set observations, skew–normal model: (a) First 5,000 iterations of the Log–posterior (R); (b) First 100,000 Log–posterior (R); (c) First 5,000 iterations of the Log–posterior (CR); (d) First 100,000 Log–posterior (CR).



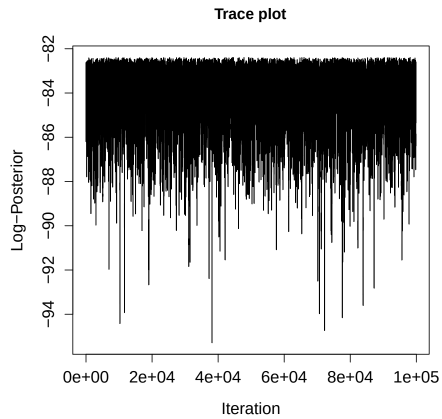
(a)



(b)



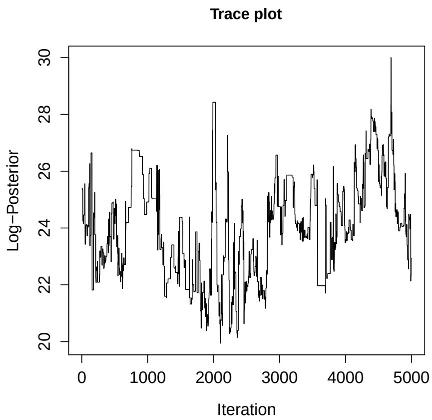
(c)



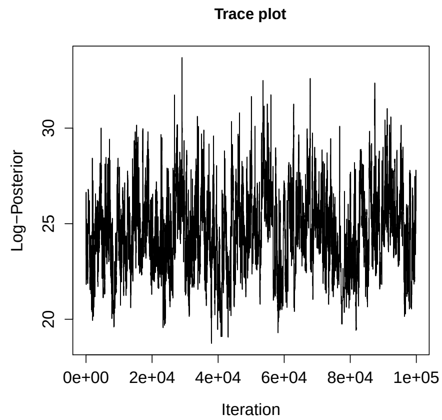
(d)

Figure G.21: Set observations, two–piece normal model: (a) First 5,000 iterations of the Log–posterior (R); (b) First 100,000 Log–posterior (R); (c) First 5,000 iterations of the Log–posterior (CR); (d) First 100,000 Log–posterior (CR).

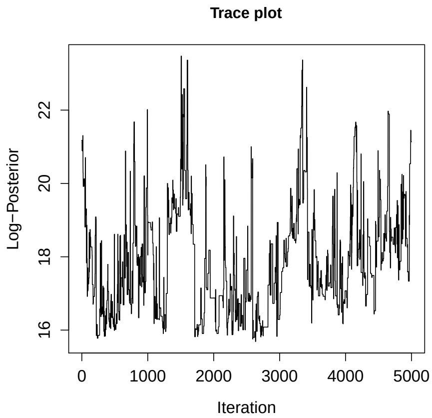
G.4 Trace plots for Chapter 6



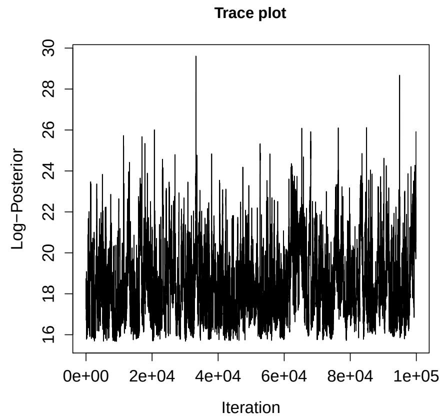
(a)



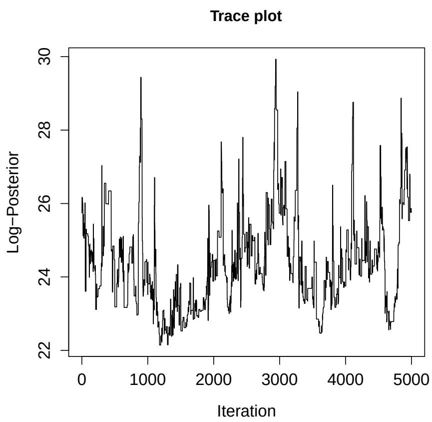
(b)



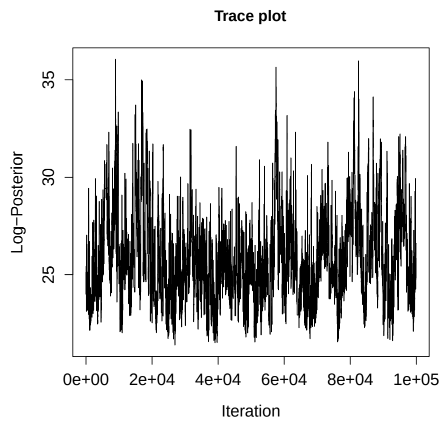
(c)



(d)



(e)



(f)

Figure G.22: Fibre glass data: (a) First 5,000 iterations of the Log–posterior (DTP); (b) First 100,000 Log–posterior (DTP); (c) First 5,000 iterations of the Log–posterior (TPSC); (d) First 100,000 Log–posterior (TPSC); (e) First 5,000 iterations of the Log–posterior (TPSH); (f) First 100,000 Log–posterior (TPSH).

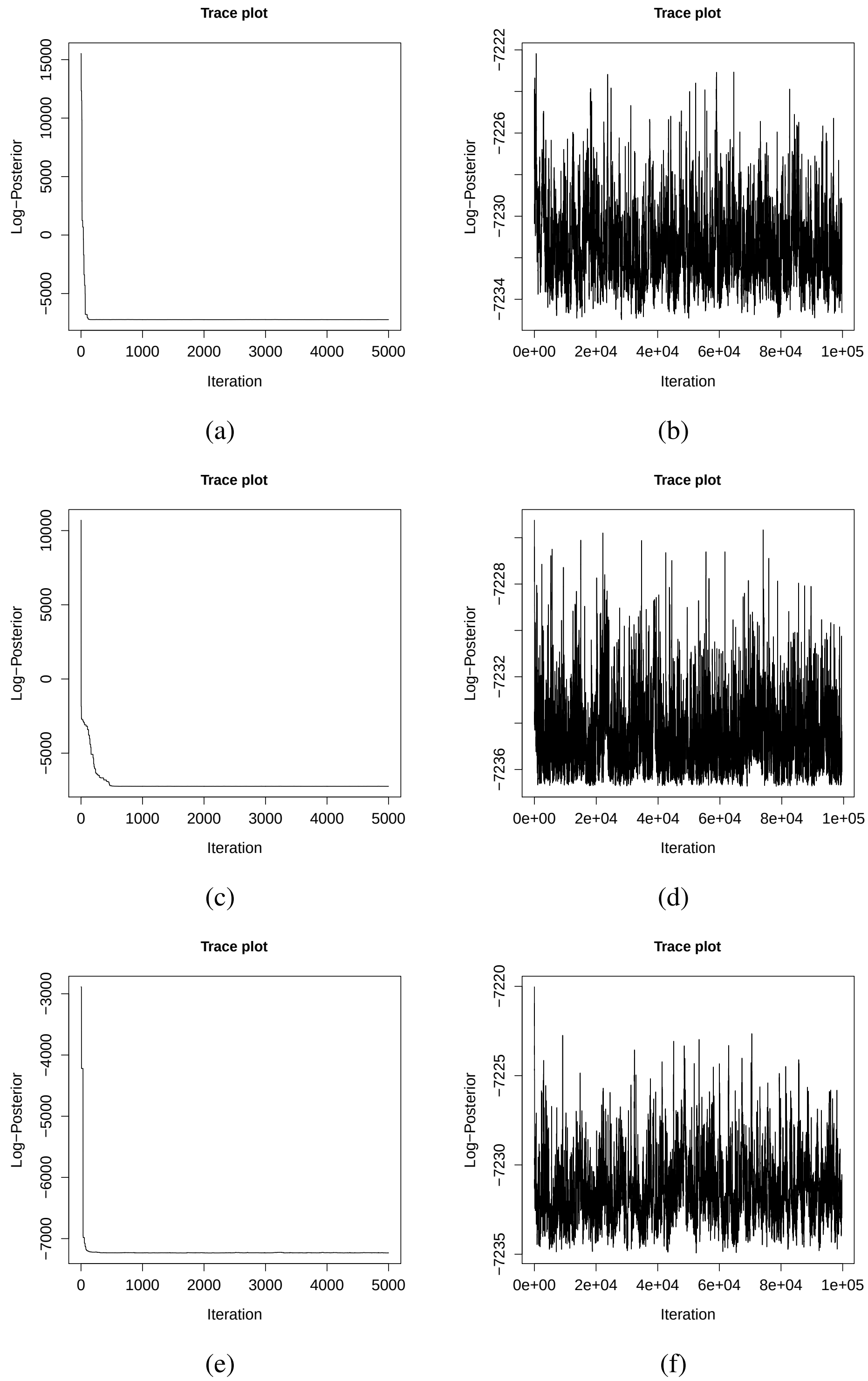


Figure G.23: EUR/NOK exchanges rates data: (a) First 5,000 iterations of the Log–posterior (DTP); (b) First 100,000 Log–posterior (DTP); (c) First 5,000 iterations of the Log–posterior (TPSC); (d) First 100,000 Log–posterior (TPSC); (e) First 5,000 iterations of the Log–posterior (TPSH); (f) First 100,000 Log–posterior (TPSH).

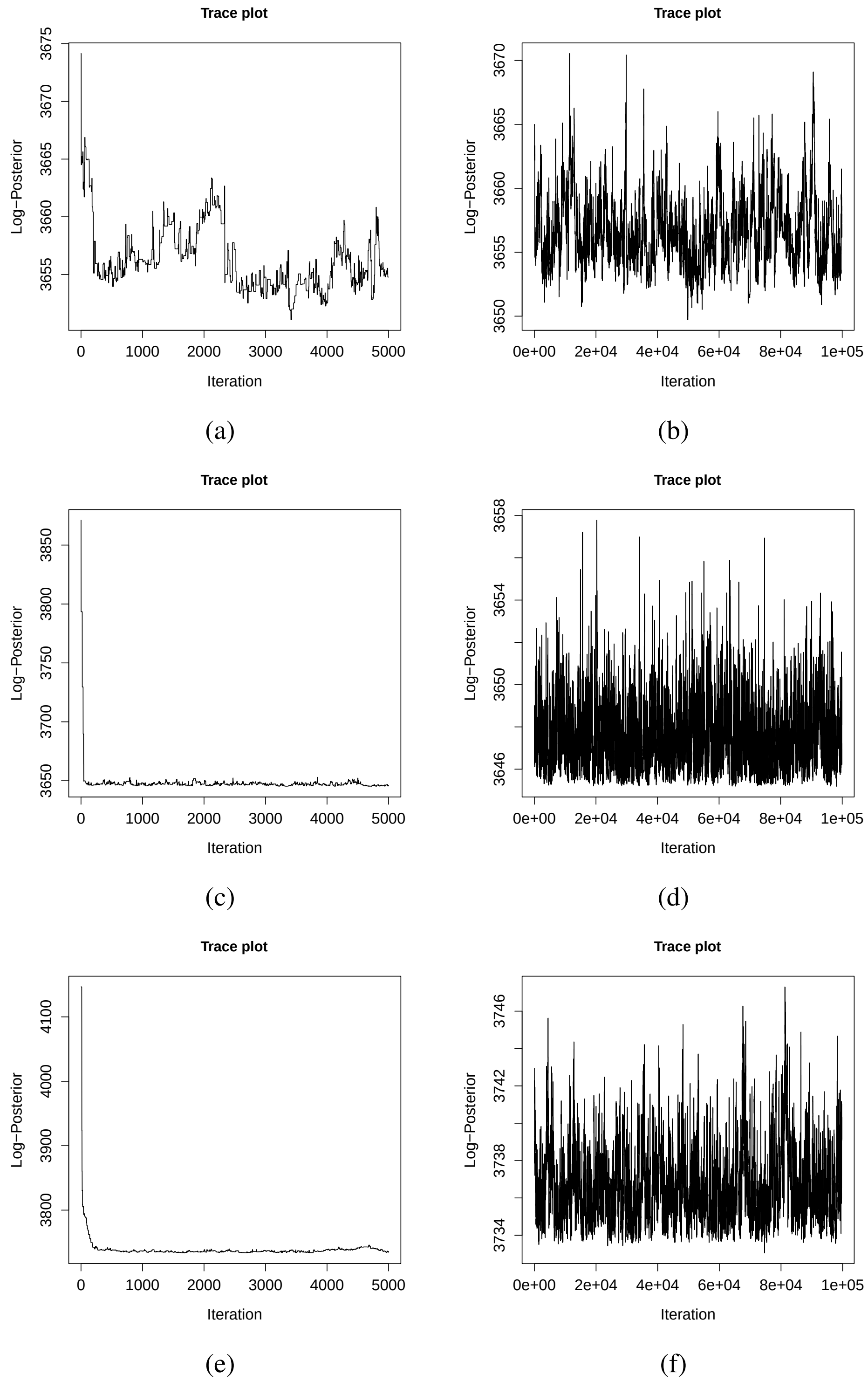


Figure G.24: Aon data: (a) First 5,000 iterations of the Log–posterior (DTP); (b) First 100,000 Log–posterior (DTP); (c) First 5,000 iterations of the Log–posterior (TPSC); (d) First 100,000 Log–posterior (TPSC); (e) First 5,000 iterations of the Log–posterior (TPSH); (f) First 100,000 Log–posterior (TPSH).

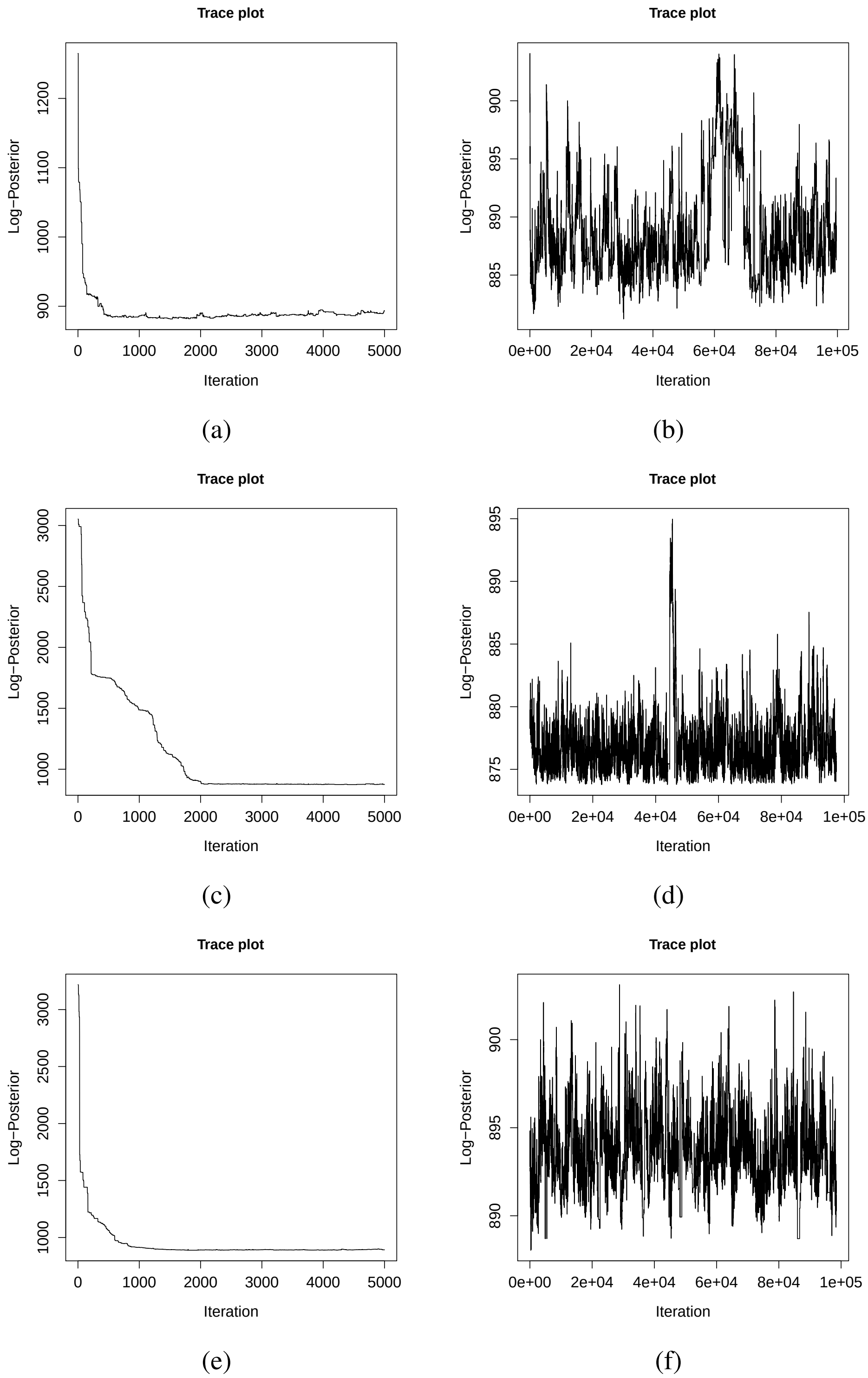


Figure G.25: Biometric measurements data: (a) First 5,000 iterations of the Log–posterior (DTP); (b) First 100,000 Log–posterior (DTP); (c) First 5,000 iterations of the Log–posterior (TPSC); (d) First 100,000 Log–posterior (TPSC); (e) First 5,000 iterations of the Log–posterior (TPSH); (f) First 100,000 Log–posterior (TPSH).

Bibliography

Aas, K. and Haff, I. H. (2006). The generalized hyperbolic skew Student's $t-$distribution. *Journal of Financial Econometrics* 4: 275–309.

Abtahi, A. and Towhidi, M. (2011). The new unified representation of multivariate skewed distributions. *Statistics: A Journal of Theoretical and Applied Statistics* DOI:10.1080/02331888.2011.577896

Arellano-Valle, R. and Azzalini A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* 33: 561–574.

Arellano-Valle, R.B., Gómez, H. W. and Quintana, F. A. (2005). Statistical inference for a general class of asymmetric distributions. *Journal of Statistical Planning and Inference* 128: 427–443.

Arnold, B. C. and Beaver, R. J. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting (with discussion), *Test* 11: 7–54.

Arnold, B. C., Beaver, R. J., Groeneveld, R. A. and Meeker, W. Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika* 58: 471-478.

Arnold, B. C. and Groeneveld, R. A. (1995). Measuring skewness with respect to the mode. *The American Statistician* 49: 34–38.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* 12: 171-178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica* 46: 199–208.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of Royal Statistical Society* Series B 65: 367–389.

Azzalini, A. and Chiogna, M. (2004). Some results on the stress-strength model for skew-normal variates. *Metron* LXII: 315-326.

Baklizi, A. and Eidous, O. (2006). Nonparametric estimation of $P(X < Y)$ using kernel methods. *Metron* LXIV: 47–60.

Balanda, K. P. and Macgillivray, H. L. (1988). Kurtosis: a critical review. *The American Statistician* 42: 111–119.

Balanda, K. P. and Macgillivray, H. L. (1990). Kurtosis and spread. *Canadian Journal of Statistics* 18: 17–30.

Barbiero, A. (2012). Interval estimators for reliability: the bivariate normal case. *Journal of Applied Statistics* 39: 501–512.

Barndorff-Nielsen, O. E. (1977). Exponentially Decreasing Distributions for the Logarithm of Particle Size. *Proceedings of the Royal Society London A*, 353: 401-419.

Bayes, C. L. and Branco, M. D. (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian Journal of Probability and Statistics* 21: 141–163.

Berger, J. O. and Sun, D. (2008). Objective priors for the bivariate normal model. *Annals of Statistics* 36: 963–982.

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *Annals of Statistics* 37: 905–938.

Berlaint, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, New York.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, Mass: Addison-Wesley.

Carta, A. and Steel, M. F. J. (2012). Modelling multi-output stochastic frontiers using copulas. *Computational Statistics and Data Analysis* 56: 3757–3773.

Chen, Q., Gerlach, R. and Lu, Z. (2011). Bayesian Value-at-Risk and expected shortfall forecasting via the asymmetric Laplace distribution. *Computational Statistics and Data Analysis* 56: 3498-3516.

Chopin, N. and Robert, C. P. (2010). Properties of nested sampling. *Biometrika* 97: 741–755.

Christen, J. A. and Fox, C. (2010). A general purpose sampling algorithm for continuous distributions (the t-walk). *Bayesian Analysis* 5: 263–282.

Clarke, B. and Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference* 41: 37–60.

Cox, D. R. and Reid, N. (1987). Orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society*, Series B 49: 1–39.

Critchley, F. and Jones, M. C. (2008). Asymmetry and gradient asymmetry functions: density-based skewness and kurtosis. *Scandinavian Journal of Statistics* 35: 415-437.

Datta, G. S. and Ghosh, J. K. (1995). Noninformative priors for maximal invariant parameter in group model. *Test* 4: 95–114.

Díaz-Francés, E. and Montoya J. A. (2012). The simplicity of likelihood based inferences for $P(X < Y)$ and for the ratio of means in the exponential model. *Statistical Papers* 54: 499–522.

Edgeworth, F. Y. (1904). The law of error. *Transactions of the Cambridge Philosophical Society* 20: 36–65 and 113–141.

Efron, B. and Gous, A. (2001). *Scales of Evidence for Models Selection: Fisher versus Jeffreys.* IMS Lecture Notes – Monograph Series, Volume 38.

Enis, P. and Geisser, S. (1971). Estimation of the probability that $Y < X$. *Journal of the American Statistical Association* 66: 162–168.

Everitt B.S. (2002). *Cambridge Dictionary of Statistics*, 2nd Edition. Cambridge University Press, Cambridge, UK.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions.* Chapman & Hall, London.

Fechner, G. T. (1897). *Kollectivmasslehre.* Leipzig, Engleman.

Fernández, C., Osiewalski, J. and Steel, M. F. J. (1995). Modeling and inference with v-spherical distributions. *Journal of the American Statistical Association* 90: 1331-1340.

Fernández, C. and Steel, M. F. J. (1998a). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* 93, 359–371.

Fernández, C. and Steel, M. F. J. (1998b). On the dangers of modelling through continuous distributions: A Bayesian perspective, in Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. eds., *Bayesian Statistics 6*, Oxford University Press (with discussion), pp. 213–238.

Fernández, C. and Steel, M. F. J. (1999a). Multivariate Student-t regression models: Pitfalls and inference. *Biometrika* 86, 153–167.

Fernández, C. and Steel, M. F. J. (1999b). Reference priors for the general location-scale model. *Statistics and Probability Letters* 43, 377–384.

Fernández, C. and Steel, M. F. J. (2000). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory* 16: 80–101.

Ferreira, J. T. A. S. and Steel, M. F. J. (2006). A constructive representation of univariate skewed distributions. *Journal of the American Statistical Association*, Theory and Methods, 101: 823–829.

Ferreira, J. T. A. S. and Steel, M. F. J. (2007). A new class of skewed multivariate distributions with applications to regression analysis. *Statistica Sinica* 17: 505–529.

Fieller, N. R. J., Flenley, E. C. and Olbricht, W. (1992). Statistics of particle size data. *Applied Statistics* 41, 127–146.

Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* 41: 933–945.

Fischer, M. J. (2004). The L-distribution and skew generalizations. Diskussionspapier 63/2004.

Fischer, M. J. (2012). A skew and leptokurtic distribution with polynomial tails and characterizing functions in closed form. *IWQW Discussion Paper series*, No. 03/2012.

Fischer, M. and Klein, I. (2004). Kurtosis modelling by means of the $J-$transformation. *Allgemeines Statistisches Archiv* 88: 35–50.

Fischer, M. J. and Vaughan, D. (2010). The Beta-Hyperbolic Secant (BHS) distribution. *Austrian Journal of Statistics* 39: 245-258.

Fonseca, T. C. O., Ferreira, M. A. R and Migon, H. S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika* 95: 325–333.

Fonseca, T. C. O., Migon, H. S. and Ferreira, M. A. R. (2012). Bayesian analysis based on the Jeffreys prior for the hyperbolic distribution. *Brazilian Journal of Probability and Statistics* 4: 327–343.

Galbraith, J.W. and van Norden, S. (2012). Assessing gross domestic product and inflation forecasts derived from Bank of England fan charts. *Journal of the Royal Statistical Society, A* 175: 713–727.

García, V.J., Gómez-Déniz, E., Vázquez-Polo, F.J. (2010). A new skew generalization of the Normal distribution: Properties and applications. *Computational Statistics and Data Analysis* 54: 2021–2034.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society*, Series B 56: 501-514.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis* 3: 515–534.

Genç, A. I. (2012). Estimation of $P(X > Y)$ with Topp-Leone distribution. *Journal of Statistical Computation and Simulation* 83: 326–339.

George, D. and George, S. (2011). Marshall-Olkin Esscher transformed Laplace distribution and processes. *Brazilian Journal of Probability and Statistics*, forthcoming

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57: 1317-1339.

Gibbons, J. F. and Mylroie, S. (1973). Estimation of impurity profiles in ion-implanted amorphous targets using joined half-Gaussian distributions. *Applied Physics Letters* 22: 568–569.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* 102: 360–378.

Goerg, G. M. (2011). Lambert W Random Variables - A New Generalized Family of Skewed Distributions with Applications to Risk Estimation. *The Annals of Applied Statistics* 5: 2197–2230.

Greco, L. and Ventura, L. (2011). Robust inference for the stress-strength reliability. *Statistical Papers* 52: 773–788.

Groeneveld, R. A. (1991). An influence function approach to describing the skewness of a distribution. *The American Statistician* 45: 97–102.

Groeneveld, R. A. (1998). A class of quantiles measures for kurtosis. *The American Statistician* 52: 325–329.

Groeneveld, R. A. and Meeden, G. (1984). Measuring Skewness and Kurtosis. *Journal of the Royal Statistical Society*, Series D 33: 391-399.

Groeneveld, R. A. and Meeden, G. (2009). An improved skewness measure. *Metron*, 67: 325–337.

Gupta, R. C. and Brown, N. (2001). Reliability studies of the skew-normal distribution and its application to a strength-stress model. *Communications in Statistics: Theory and Methods* 30: 2427-2445.

Gupta, R. D. and Gupta, R. C. (2008). Analyzing skewed data by power normal model. *Test* 17: 197–210.

Gupta, R. C. and Peng, C. (2009). Estimating reliability in proportional odds ratio models. *Computational Statistics & Data Analysis* 53: 1495–1510.

Haynes, M. A., MacGilllivray, H. L. and Mergersen, K. L. (1997). Robustness of ranking and selection rules using generalized g and k distributions. *Journal of Statistical Planning and Inference* 65: 45–66.

Heinz G., Peterson, L. J., Johnson, R. W. and Kerk, C. J. (2003). Exploring relationships in body dimensions. *Journal of Statistics Education* (online only) 11(2). www.amstat.org/publications/jse/v11n2/datasets.heinz.html.

Heitjan, D. F. (1989). Inference from grouped continuous data: A review. *Statistical Science* 4: 164-183.

Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analyisis of generalized linear models using Jeffreys's priors. *Journal of the American Statistical Association* 86: 981–986.

Jarner, S. F. and Roberts, G. O. (2007). Convergence of Heavy–tailed Monte Carlo Markov Chain Algorithms. *Scandinavian Journal of Statistics* 34: 781-815.

Jeffreys, H. (1941). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London*: Series A, 183: 453–461.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.) Oxford: Clarendon.

Jing, B. Y., Yuan, J. and Zhou, W. (2009). Jackknife empirical likelihood. *Journal of the American Statistical Association* 104, 1224–1232.

John, S. (1982). The three-parameter two-piece normal family of distributions and its fitting. *Communications in Statistics - Theory and Methods* 11: 879-885.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, Vol. 2 (2nd Ed. ed.). John Wiley & Sons, Inc., New York.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika* 36: 149-176.

Jones, M. C. (2004). Families of distributions arising from distributions of order statistics. *TEST* 13: 1–43.

Jones, M. C. (2006). A note on rescalings, reparametrizations and classes of distributions. *Journal of Statistical Planning and Inference* 136: 3730-3733.

Jones, M. C. and Anaya-Izquierdo K. (2010). On parameter orthogonality in symmetric and skew models. *Journal of Statistical Planning and Inference* 141: 758–770.

Jones, M. C. and Faddy, M. J. (2003). A skew extension of the t-distribution, with applications. *Journal of Royal Statistical Society* Series B 65: 159–174.

Jones, M. C. and Pewsey A. (2009). Sinh-arcsinh distributions. *Biometrika* 96: 761–780.

Juárez, M. A. and Steel, M. F. J. (2010). Model-based clustering of non-Gaussian panel data based on skew-t distributions. *Journal of Business and Economic Statistics* 28: 52–66.

Julià, O. and Vives-Rego, J. (2005). Skew-Laplace distribution in Gram-negative bacterial axenic cultures: new insights into intrinsic cellular heterogeneity. *Microbiology* 151: 749–755.

Julià, O. and Vives-Rego, J. (2008). A microbiology application of the skew-Laplace distribution. *SORT* 32: 141–150.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* 90, 773–795.

Klein, I. and Fischer, M. (2006a). Power kurtosis transformations: Definition, properties and ordering. *Allgemeines Statistisches Archiv* 90: 395–401.

Klein, I. and Fischer, M. (2006b). Skewness by splitting the scale parameter. *Communications in Statistics - Theory and Methods* 90: 395–401.

Kotz, S., Kozubowski, T. J. and Podgorski, K. (2001). *The Laplace distribution and generalizations.* Birkhäuser, Berlin.

Kotz, S., Lumelskii, S. and Pensky, M. (2003). *The Stress-Strength Model and its Generalizations.* Theory and Applications. Singapore: World Scientific.

Lachos, V. H., Ghosh, P. and Arellano-Valle, R. (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica* 20: 303–322.

Lehmann, E. L. (1953). The power or rank tests. *The Annals of Mathematical Statistics* 24: 23–43.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation.* Springer, New York.

Ley, C. and Paindaveine, D. (2010a). Multivariate skewing mechanisms: A unified perspective based on the transformation approach. *Statistics* & *Probability Letters* 80: 1685-1694.

Ley, C. and Paindaveine, D. (2010b). On Fisher Information Matrices and Profile Log-Likelihood Functions in Generalized Skew-Elliptical Models. *METRON* International Journal of Statistics 68: 235–250.

Ley, C. and Paindaveine, D. (2010). On the singularity of multivariate skew-symmetric models. *Journal of Multivariate Analysis* 101: 1434–1444.

Lingappaiah, G. S. (1988). On two-piece double exponentail distributions. *Journal of the Korean Statistical Society* 17, 46–55.

Liseo, B. and Loperfido, N. (2006). A note on reference priors for the scalar skew-normal distribution. *Journal of Statistical Planning and Inference* 136: 373-389.

Maiti, S. S. and Dey, M. (2012). Tilted normal distribution and its survival properties. *Journal of Data Science* 10: 225–240.

Marshall A. W., Olkin I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika* 84: 641–652.

Martinez J. and Iglewicz, B. (1984) Some properties of the Tukey g and h family of distributions. *Communications in Statistics: Theory and Methods* 13: 353–369.

Mengersen, K. L., Pudlo, P. and Robert, C. P. (2012). Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences of the United States of America* 110: 1321–1326.

Montoya, J. A. (2008). *La verosimilitud perfil en la Inferencia Estadística.* PhD Thesis, Centro de Investigación en Matemáticas A. C., México.

Mudholkar, G. S. and Hutson, A. D. (2000). The epsilon-skew-normal distribution for analyzing near-normal data. *Journal of Statistical Planning and Inference* 83: 291–309.

Murillo A. and Rubio, F. J. (2011). A note on the infinite divisibility of a class of transformations of normal variables. *Brazilian Journal of Probability and Statistics*, forthcoming.

Nadarajah, S. (2005). Reliability for some bivariate gamma distributions. *Mathematical Problems in Engineering* 2005: 151–163.

Nadarajah, S. (2006). The exponentiated Gumbel distribution with climate application. *Environmetrics* 17: 13-23.

Nandi, S. B. and Aich, A. B. (1994). A note on confidence bounds for $P(X > Y)$ in bivariate normal samples. *Sankhyā, Ser. B* 56: 129–136.

Oja, H. (1981). On location, scale, skewness and kurtosis of univariate distributions. *Scandinavian Journal of Statistics* 8: 154–168.

Pearson, K. (1895). Contributions to the mathematical theory of evolution, II: skew variation in homogeneous material. *Transactions of the Royal Philosophical Society*, Ser. A 186: 343-414.

Pearson, K. (1905). Skew variation, a rejoinder. *Biometrika* 4: 169–212.

Pewsey, A. (2000). Problems of inference for Azzalini's skew-normal distribution. *Journal of Applied Statistics* 27: 859–870.

Pinheiro, M. (2012). Marginal distributions of random vectors generated by affine transformations of independent two-piece normal variables. *Journal of Probability and Statistics* vol. 2012, Article ID 758975, 10 pages. doi:10.1155/2012/758975

Punathumparambath B., Kulathinal, S. and George, S. (2012). Asymmetric type II compound Laplace distribution and its application to microarray gene expression. *Computational Statistics and Data Analysis* 56: 1396–1404.

Purdom E., Holmes S.P. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology* 4: Article 16.

Ramirez-Cobo, P., Lillo, R. E., Wilson, S. and Wiper, M. P. (2010). Bayesian inference for double Pareto lognormal queues. *The Annals of Applied Statistics* 4: 1533–1557.

Rayner, G. D. and MacGillivray (2002). Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing* 12: 57-75.

Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*(2nd ed.). New York: Springer.

Romano, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics* 16: 629–647.

Rosco, J. F., Jones, M. C. and Pewsey, A. (2011). Skew t distributions via the sinh-arcsinh transformation. *TEST* 30: 630–652.

Rubio, F. J. and Steel, M. F. J. (2011a). Inference for grouped data with a truncated skew-Laplace distribution. *Computational Statistics and Data Analysis* 55: 3218-3231.

Rubio, F. J. and Steel, M. F. J. (2011b). Inference in Two-Piece Location-Scale models with Jeffreys Priors. CRiSM working paper 11–13.

Rubio, F. J. and Steel, M. F. J. (2012). On the Marshall-Olkin transformation as a skewing mechanism. *Computational Statistics and Data Analysis* 56: 2251-2257.

Rubio, F. J. and Steel, M. F. J. (2013). Bayesian inference for $P(X < Y)$ using asymmetric dependent distributions. *Bayesian Analysis* 8: 43–62.

Sahu, S. J., Dey, D. K. and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics* 2: 129–150.

Schneeweiss, H., Komlos, J. and Ahmad, A.S. (2010). Symmetric and asymmetric rounding: A review and some new results. *Advances in Statistical Analysis* 94: 247–271.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics* 6: 461-464.

Seaman III, J. W., Seaman Jr., J. W. and Stamey, J. D. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician* 66: 77–84.

Shoukri, M. M., Chaudary, M. A. and Al-Halees, A. (2005). Estimating $P(Y < X)$ when X and Y are paired exponential variables. *Journal of Statistical Computation and Simulation* 75: 25-38.

Smith, R. L. and Naylor, J. C. (1987). A Comparison of Maximum Likelihood and Bayesian Estimators for the Three- Parameter Weibull Distribution. *Journal of the Royal Statistical Society*, Series C 36: 358–369.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, Series B 64: 583-639.

Sun, D., Ghosh, M. and Basu, A. P. (1998). Bayesian analysis for a stress–strength system under noninformative priors. *The Canadian Journal of Statistics* 26: 323–332.

Taylor, J. M. G., Siqueira, A. L. and Weiss, R. T. (2000). The cost of adding parameters to a model. *Journal of the Royal Statistical Society*, Series B 58: 593–607.

Trindade, A.A. and Zhu, Y. (2007). Approximating the distributions of estimators of financial risk under an asymmetric Laplace law. *Computational Statistics and Data Analysis* 51: 3433–3447.

Trindade, A. A., Zhu, Y. and Andrews, B. (2010). Time series models with asymmetric Laplace innovations. *Journal of Statistical Computation and Simulation* 80: 1317–1333.

Tukey, J. M. (1960). The practical relationship between the common transformations of counts of amounts. *Princeton University Statistical Techniques Research Group*, Technical Report No. 36, Princeton.

Tukey, J. M. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, M. A.

van Zwet, W. R. (1964). *Convex Transformations of Random Variables.* Mathematisch Centrum, Amsterdam.

Venkatraman, E. S. and Begg, C. B. (1996). A distribution-free procedure for comparing operating characteristic curves from a paired experiment. *Biometrika* 83: 835–848.

Ventura, L. and Racugno, W. (2011). Recent advances on Bayesian inference for $P(X < Y)$. *Bayesian Analysis* 6: 1–18.

Venturini, S., Dominici, F. and Parmigiani, G. (2008). Gamma shape mixtures for heavy-tailed distributions. *Annals of Applied Statistics* 2: 756–776.

Wagner, L. A. (2007). Some skew models for quantal response analysis. University of Rochester, PhD thesis.

Wallis, K. (2004). An assessment of Bank of England and National Institute inflation forecast uncertainties. *National Institute Economic Review* 189: 64–71.

Wallis, K. (2013). The two-piece normal, binormal, or double Gaussian distribution: its origin and rediscoveries. *Statistical Science*, forthcoming.

Wang, J., Boyer, J. and Genton M. C. (2004). A skew symmetric representation of multivariate distributions. *Statistica Sinica* 14: 1259–1270.

Zelterman, D. (1987). Parameter estimation in the generalized logistic distribution. Computational Statistics & Data Analysis 5: 177–184.

Zhou, W. (2008). Statistical inference for $P(X < Y)$. *Statistics in Medicine* 27: 257–279.

Zhu, D. and Galbraith, J. W. (2010). A generalized asymmetric Student-t distribution with application to financial econometrics. *Journal of Econometrics* 157: 297–305.

Zhu, D. and Galbraith, J. W. (2011). Modeling and forecasting expected shortfall with the generalized asymmetric Student-t and asymmetric exponential power distributions. *Journal of Empirical Finance* 18: 765–778.